FAIRNESS IN COLLEGE ADMISSION EXAMS: FROM TEST SCORE GAPS TO EARNINGS INEQUALITY

EVAN RIEHL*
CORNELL UNIVERSITY

JUNE 2019

Abstract. This paper asks whether reducing socioeconomic gaps in college admission exam scores also reduces earnings inequality. A simple framework shows that the answer to this question depends on the exam's predictive power for students' earnings returns to college quality. I exploit a redesign of the Colombian national college admission exam to estimate these returns. Low-income students who took the new exam received higher scores and attended better colleges, but they had *lower* earnings after college. Thus exams with lower score gaps may not reduce earnings inequality if they are poor predictors of which students would succeed at better colleges.

* Department of Economics, Cornell University, 266 Ives Hall, Ithaca, NY 14853 (e-mail: eriehl@cornell.edu).
A previous version of this paper was titled "Assortative Matching and Complementarity in College Markets." For useful comments I thank Peter Arcidiacono, Peter Bergman, Serena Canaan, Nicolás De Roux, Christian Dustmann, Christopher Hansman, Adam Kapor, Michael Lovenheim, W. Bentley MacLeod, Costas Meghir, Jonah Rockoff, Miikka Rokkanen, Juan E. Saavedra, Judith Scott-Clayton, Mallika Thomas, Miguel Urquiola, and Eric Verhoogen. I am grateful to Luis Omar Herrera Prada for invaluable help with the data. All errors are my own.

Do standardized college admission exams reduce or exacerbate inequality? The debate over admission exams usually centers on how they affect access to selective colleges. These exams aim to reduce the influence of family wealth in admissions, and yet most tests find large achievement gaps between socioeconomic groups. This fact leads some to argue that admission exams are biased against low-income students, and it has motivated a growing number of U.S. colleges to adopt test-optional admission policies.[1]

This paper takes a different approach to this debate by asking how college admission exams affect inequality in labor market outcomes. I first develop a framework that relates socioeconomic gaps in admission exam scores to inequality in post-college earnings. This link depends on how well the exam predicts students' *returns to college quality*—their potential earnings benefits to attending a better school. If returns are heterogeneous, exams with low test score gaps but poor predictive power may cause "mismatch" in admissions (Arcidiacono and Lovenheim, 2016), potentially harming low-income students in the labor market.

I then estimate heterogeneity in returns to college quality using data and a natural experiment from the country of Colombia. The data link admission scores, college choices, and earnings one decade later for nearly all students in the country. For identification, I exploit an overhaul of the national admission exam that dramatically reduced test score gaps between socioeconomic groups and shifted low-income students into better colleges. I find substantial heterogeneity in returns to college quality including negative returns for some low-income students. The new exam was a worse predictor of these returns, and it led to lower earnings for *both* high- and low-income students relative to previous cohorts. These results document mismatch under weaker assumptions than those in existing work (Arcidiacono et al., 2016), and they show that lower test score gaps do not necessarily reduce earnings inequality.

The paper is organized as follows. Section 1 develops a model that shows how socioeconomic test score gaps affect post-college earnings. Students are defined by their socioeconomic status (SES) and a vector of abilities that reflect capacity to answer potential exam questions. A testing agency chooses weights on each ability, which yields an exam score for each student. The agency can reduce test score gaps by measuring abilities that are less correlated with SES, but this may affect the exam's predictive power for student outcomes.

The key model parameter is the return to college quality. I assume students sort into colleges by their admission exam scores, and define a college's quality as the mean test score of its student body (Dale and Krueger, 2002; Hoxby, 2009; MacLeod et al., 2017). The return to college quality, which I denote by ${}_i$, is the causal effect of college quality on student i's earnings. Importantly, these returns can differ across individuals.

The labor market implications of the admission test depend on the distribution of returns to college quality and the exam's predictive power for these returns. If ${}_i$ is positive and

[1] For example, Colby College and the University of Chicago switched to test-optional admissions in 2018.

constant across individuals, then reducing test score gaps lowers earnings inequality without changing average earnings in the market. If instead $_i$ is correlated with SES, there can be an equity/efficiency tradeoff in reducing test score gaps.

Reducing test score gaps can also lead to *lower* earnings for low SES students in some cases. If the average value of $_i$ is negative for low SES students, reducing test score gaps shifts some students into better colleges but lowers their earnings returns. This is a version of the "mismatch hypothesis," which states that some students may be better off at lower ranked colleges. Low SES earnings can also fall if the test becomes a worse predictor of heterogeneity in $_i$. In this case, there is "mismatch" because low SES students with high returns to college quality are less likely to gain admission to top colleges.

The framework shows that it is important for admission exams to predict students' potential returns to college quality. In the rest of the paper, I exploit a natural experiment to estimate heterogeneity in these returns. In 2000, the Colombian national college admission exam underwent a major overhaul with the goal of reducing socioeconomic bias. The new exam tested "competencies" rather than "content," and was similar in spirit to the 2016 redesign of the U.S. SAT. I measure the reform's effects on students' test scores, college quality, and earnings one decade later using individual-level administrative data. I then combine these results to characterize heterogeneity in returns to college quality.

In Section 2, I show that the reform dramatically reduced socioeconomic test score gaps, but the exam became a worse predictor of students' college outcomes. For example, the test score gap between students in the top and bottom income quartiles fell by 50 percent in several subjects. But the admission exam's validity—measured by its correlation with college exit exam scores and graduation rates—declined by a similar magnitude.

Section 3 describes my identification strategy for longer-run effects of the reform. In Colombia, most students attend college close to home, and many low-income students can only afford to attend public universities. At most public universities, admission is determined solely by national exam scores, but some schools administer a separate entrance exam. These different admission methods created geographic variation in the stakes of the national exam reform. I define students as "treated" by the reform if they lived near public universities that used national exam scores for admissions. "Control" students lived near schools with other admission criteria. My empirical specification is a differences-in-differences model that estimates changes in outcomes across exam cohorts and treated/control areas.

Section 4 shows how the reform affected college and earnings outcomes. In treated areas, low SES students attended better colleges after the reform, and high SES students were displaced to lower-quality schools. In net the reform reduced the SES "college quality gap" by ten percent. Relative to earlier cohorts, however, the reform led to a one percent decrease in earnings for *both* high and low SES students. Graduation rates also declined for both groups, showing that the reform caused some students to drop out of college. These results are robust to several definitions of reform treatment, and I show that outcomes fell sharply in the first cohort of the new exam. I present evidence that results are not driven by changes in concurrent economic conditions or overall college enrollment rates.

Section 5 combines the college and earnings effects to estimate heterogeneity in returns to college quality. I adapt the differences-in-differences model into a two-stage least squares regression that uses a student's exam cohort and place of residence as an instrument for college quality. I then use new methods for heterogeneous causal effects (Athey et al., 2017) to estimate variation in returns using a large vector of individual covariates.

I find substantial heterogeneity in returns to college quality, and show that the post-reform exam had less predictive power for these returns. Among students affected by the reform, returns to college quality were strongly correlated with SES and were *negative* for many low SES students. On average, a ten percentile point increase in college quality caused a nine percent decrease in low SES earnings, but there is significant variation in these returns. I show that the reform reduced the correlation between admission scores and returns to college quality, with a decline in predictive power both across and within SES groups.

In sum, the Colombian reform reduced low SES students' earnings for two reasons. First, some students who were induced to attend better schools had negative returns to college quality. Second, the new exam became a worse predictor of these returns, making it less likely that high-return students were admitted to top colleges. More broadly, reducing test score gaps—or eliminating admission exams altogether—may help low SES students only if there are other ways to identify which students are likely to succeed at top colleges.

This paper relates most directly to work on college mismatch. Prior work has shown that affirmative action can, in some cases, reduce graduation rates for minority students (Arcidiacono et al., 2011, 2014, 2016), but these papers make strong assumptions about unobservable determinants of college choice.[2] My results rely only on a parallel trends assumption, and I document mismatch in both graduation rates and earnings. Further, I show that mismatch can arise from either negative or heterogenous returns to college quality (Andrews et al., 2016; Dillon and Smith, 2018), which builds on work on heterogeneity in the return to schooling (Card, 2001; Carneiro et al., 2003, 2011). An important caveat is that socioeconomic mismatch need not imply mismatch from race-based affirmative action, and may be more likely when test scores are the only admission criterion.

My paper helps reconcile evidence of mismatch with work on the returns to college selectivity (Dale and Krueger, 2002; Hoekstra, 2009; Saavedra, 2009; Hastings et al., 2013; Canaan

[2] Other work exploits affirmative actions bans but finds no direct evidence of mismatch (Cortes, 2010; Backes, 2012; Hinrichs, 2012, 2014). A related literature examines the causes of "overmatch" or "undermatch" in college admissions (Hoxby and Avery, 2013; Smith et al., 2013; Dillon and Smith, 2017).

and Mouganie, 2015; Kirkebøen et al., 2016; Chetty et al., 2017; Hoxby, 2018; Zimmerman, 2018). This work uses well-identified research designs and often finds that disadvantaged students have positive returns to selective colleges or majors. Estimates can vary with the empirical strategy if returns to college quality are heterogeneous. Many of these papers identify returns for students who are marginally admitted to selective programs, but in mismatch work—as in my context—large-scale policies cause students to attend colleges where they otherwise would not have gained admission. Thus mismatch may only arise when students are much less prepared than their college peers.

Lastly, this paper relates to work on the design of college admission exams (Rothstein, 2004; Bettinger et al., 2013; Hoxby and Turner, 2013; Bulman, 2015; Goodman, 2016; Goodman et al., 2018).[3] This work shows how exams affect college access, while my paper explores their labor market implications. Exams that improve access for low-income students may not raise their earnings if they are poor predictors of the ability to succeed at top colleges.

1. A model of admission exams and post-college earnings

This section develops a framework to show how socioeconomic gaps in college admission test scores affect the distribution of post-college earnings. I first describe how admission exam design affects score gaps and predictive power for student outcomes. I then define the return to college quality—the parameter that connects the markets for college admissions and post-college labor. Lastly, I show that the link from test scores to earnings inequality depends on how well the exam predicts heterogeneity in returns to college quality.

1.1. **Students.** The market for college admissions consists of a large population of students with two characteristics of interest: socioeconomic status and ability. Let X_i denote individual i's socioeconomic status (SES) as measured, for example, by parental income. For simplicity, I define X_i to be a binary indicator of high SES with $Pr[X_i = 0] = Pr[X_i = 1] = 0.5$.

I define ability as a K-dimensional vector $A_i = \{a_{1i}, \dots, a_{Ki}\}$, where each a_{ki} is an indicator for whether student i can correctly answer an exam question of type k. For example, if question k is "Solve for z in $z^2 + 2z + 1 = 0$," then $a_{ki} = 1$ for students who can find "$z = -1$" and $a_{ki} = 0$ for those who cannot. Thus A_i is a high-dimensional vector, and the a_k's can vary in their correlations with SES and with students' potential outcomes.

[3] Other matching mechanisms in education include centralized assignment (Abdulkadiroğlu et al., 2005), affirmative action (Durlauf, 2008; Bertrand et al., 2010; Bagde et al., 2016), and "percent plans" (Long, 2004; Kain et al., 2005; Niu and Tienda, 2010; Cullen et al., 2013; Daugherty et al., 2014; Kapor, 2015).

1.2. **College admission exam.** A testing agency designs a college admission exam that yields a raw score T_i^* for each student, where

$$T_i^* = A_i' w + e_i = \sum_{k=1}^{K} w_k a_{ki} + e_i.$$

The role of the testing agency is to choose a K-dimensional vector of weights $w = \{w_1, \ldots, w_K\}$ with each $w_k \in [0,1]$ and $\sum_{k=1}^{K} w_k = 1$. The error term $e_i \sim N(0, \sigma_e^2)$ is random noise reflecting factors out of the testing agency's control such as guessing.

Let $T_i = (T_i^* - \bar{T}^*)/\sigma_T$ be the normalized score, where $\bar{T}^*$ and σ_T^2 are the raw score mean and variance. If there are many questions then T_i is distributed approximately $N(0,1)$.

1.3. **Test score gaps and validity.** The testing agency can reduce test score gaps by measuring abilities that are less correlated with SES. I define the test score gap as the difference in mean exam scores between high and low SES students. Letting $\bar{T}_x = E[T_i|X_i = x]$ denote the average normalized score for SES group $x \in \{0,1\}$, the test score gap is

$$\bar{T}_1 - \bar{T}_0 = \frac{\sum_{k=1}^{K} w_k(\bar{a}_{1k} - \bar{a}_{0k})}{\sigma_T}, \tag{1}$$

where $\bar{a}_{xk}$ is the mean of the k^{th} ability for SES group x. $\bar{T}_1 - \bar{T}_0$ measures the gap between high and low SES scores in standard deviation units. This gap is smaller in exams that place more weight, w_k, on abilities with lower mean differences between SES groups, $\bar{a}_{1k} - \bar{a}_{0k}$.

The abilities measured by an exam can also affect its predictive power. Testing agencies typically measure an exam's validity by correlating test scores with college outcomes. Let Y_i be a measure of college success such as GPA or graduation, and let σ_Y^2 denote its variance. An exam's "raw validity" (Rothstein, 2004) is the correlation coefficient between Y_i and T_i:

$$\frac{\text{Cov}(Y_i, T_i)}{\sigma_Y} = \frac{\sum_{k=1}^{K} w_k \text{Cov}(Y_i, a_{ki})}{\sigma_T \sigma_Y}. \tag{2}$$

Validity increases when exams test abilities, a_{ki}, that are more related to student outcomes.

There is a potential tradeoff between test score gaps and exam validity if ability is correlated with both SES and student outcomes. A testing agency could reduce score gaps simply by increasing the relative importance of randomness in the exam, but this would also reduce the exam's predictive power. For example, an "easy" exam with questions that most students can answer ($\bar{a}_{1k} \approx \bar{a}_{0k} \approx 1$) has a low test score gap but also low validity, as most of the variation in scores comes from noise, e_i.[4] It is harder to design a test of abilities that are both weakly correlated with SES and highly predictive of student outcomes.

[4] Increasing weight on easy questions raises the noise-to-signal ratio, σ_e^2/σ_T^2. To see this, note that $\sigma_T^2 = \sum_{k=1}^{K} w_k^2 \text{ Var}(a_{ki}) + 2\sum_{j=1}^{K-1}\sum_{k=j+1}^{K} w_j w_k \text{ Cov}(a_{ji}, a_{ki}) + \sigma_e^2$. Abilities a_{ki} with $Pr[a_{ki} \approx 1]$ have a low variance and a low covariance with other abilities, a_{ji}. Thus σ_e^2 is more important in exams with more weight, w_k, on easy questions. This is also true in exams with more weight on hard questions, $Pr[a_{ki} \approx 0]$.

1.4. **Colleges.** The higher education market contains a large population of colleges with varying quality. I index colleges by c and let Q_c be a measure of their quality. In this paper I define a college's quality as the average admission exam score of its student body,

$$Q_c = E[T_i | m(i) = c], \tag{3}$$

where $m(i) = c$ is the match function that assigns student i to college c. This definition of quality is common in research on the return to college selectivity (e.g., Dale and Krueger, 2002; Hastings et al., 2013). It is also a good measure of student preferences in settings where colleges use admission scores to select students (Hoxby, 2009; MacLeod et al., 2017).

I make two assumptions on college admissions to simplify the model. First, all students have college preferences given by Q_c; any college with a larger value of Q_c is strictly preferred to another college by all students. Second, colleges are perfectly selective, so that each college admits students with a single test score value (MacLeod and Urquiola, 2015). Under these assumptions, the quality of a student's college is equal to her own exam score:

$$Q_{m(i)} = T_i. \tag{4}$$

Since T_i is a normalized score, college quality is also measured in standard deviation units. Further, Q_c is a fixed characteristic because colleges enroll students with the same scores regardless of the exam design.

My measure of college quality may not capture all ways in which the match between schools and students can affect outcomes. In particular, Q_c is likely to be a better measure of school characteristics that are fixed in the short run—such as financial resources—than traits that may change with an admission exam reform—such as peer composition. Below I discuss the potential influence of peer effects and other variable traits on my empirical results. Nonetheless, this definition of college quality is a tractable benchmark that reflects how testing agencies influence admissions in practice: by ranking individuals on their propensity for college success irrespective of student and school responses.

1.5. **Earnings and the return to college quality.** Let w_{ic} denote the log post-college earnings of student i who attends college c, and assume it is given by:

$$w_{ic} = {}_i + {}_i Q_c + \epsilon_{ic}. \tag{5}$$

Equation (5) defines earnings as a function of an individual-specific intercept, ${}_i$, college quality, Q_c, and an error term, ϵ_{ic}, that I assume satisfies $E[\epsilon_{ic}] = E[\epsilon_{ic}|X_i] = 0$.

The key parameter in the earnings equation is the coefficient on college quality, ${}_i$, which I call the *return to college quality*. This coefficient measures the percentage change in earnings from a one standard deviation increase in Q_c. This definition follows other work on the earnings impact of college selectivity (Dale and Krueger, 2002), but the i subscript allows for heterogeneity in returns across individuals. The return to college quality reflects a student's suitability for attending a higher quality college as defined by labor market outcomes. Although returns may differ at each college, $_i$ can be thought of as a linear approximation to the relationship between earnings and school quality as measured by test scores. This is a parameter of direct relevance to college admission testing agencies.

To relate exam design to earnings, let $\bar{w}_x = E[w_i|X_i = x]$ denote mean log earnings for SES group x, and use assumption (4) in equation (5). It follows that

$$\bar{w}_x = \bar{}_x + \bar{}_x\bar{T}_x + \text{Cov}(_i, T_i|X_i = x), \tag{6}$$

where $\bar{}_x$ is the mean intercept for group x, and $\bar{}_x$ is the mean return to college quality.[5] Equation (6) decomposes earnings for SES group x into a component that does not depend on the admission exam, $\bar{}_x$, a mean test score effect, $\bar{}_x\bar{T}_x$, and the within-SES covariance of test scores and returns, $\text{Cov}(_i, T_i|X_i = x)$.

1.6. **Test score gaps and earnings outcomes.** The link between test score gaps and earnings outcomes depends on the exam's predictive power for returns to college quality. To illustrate this, Proposition 1 describes the earnings effects of reducing the SES test score gap under different assumptions on the distribution of $_i$. I consider effects on three outcomes that can be defined from equation (6): mean earnings, mean low SES earnings, and earnings inequality.[6] Below I state the proposition and describe the main intuition. Appendix 1 contains a full derivation of the results.

Proposition 1. *Consider an exam reform that reduces the SES test score gap, $\bar{T}_1 - \bar{T}_0$:*

- *(A)* ***Constant returns.*** *If the return to college quality is positive and constant for all students ($_i = \ > 0$), mean earnings is unaffected and earnings inequality decreases.*
- *(B)* ***Complementarity.*** *If returns are larger for high SES students on average ($\bar{}_1 > \bar{}_0$), mean earnings decreases unless the exam becomes a better predictor of returns within SES groups, i.e.,* $\text{Cov}(_i, T_i|X_i = x)$ *increases.*
- *(C)* ***Mismatch.*** *Mean earnings for low SES students can* <u>*decrease*</u> *if either:*
  - *The mean return to college quality is negative for low SES students ($\bar{}_0 < 0$); or*
  - *The exam becomes a worse predictor of low SES returns, i.e.,* $\text{Cov}(_i, T_i|X_i = 0)$ *decreases.*

Parts (A) and (B) illustrate the potential for an efficiency/equity tradeoff in reducing SES test score gaps. When returns to college quality are homogeneous, lower test score gaps reduce earnings inequality without an efficiency cost because all students get the same

[5] Equation (6) follows from $E[_iQ_c|X_i = x] = E[_iT_i|X_i = x] = \bar{}_x\bar{T}_x + \text{Cov}(_i, T_i|X_i = x)$.

[6] Since SES is binary with $Pr[X_i = 0] = Pr[X_i = 1]$, mean earnings is $(\bar{w}_1 + \bar{w}_0)/2$ and earnings inequality is $\bar{w}_1 - \bar{w}_0$, where $\bar{w}_0$ and $\bar{w}_1$ are mean earnings for low and high SES students (equation (6)).

benefit from attending a better college. If instead returns and SES are complementary, the earnings loss from shifting high SES students into lower quality colleges may be larger than the earnings gain to the low SES students who replace them. In this case inequality may decrease, but at the cost of a decline in overall earnings.

Part (C) describes two cases in which reducing test score gaps can actually harm low SES students on the labor market. First, low SES earnings fall if their average return to college quality is negative. This condition is often called the "mismatch hypothesis" (Arcidiacono and Lovenheim, 2016), which states that some students may be better off attending lower-ranked schools if they are academically unprepared for top colleges. Second, low SES students' earnings can fall if the exam becomes a worse predictor of their returns to college quality. A reduction in predictive power for $_i$ is not mismatch in the sense that students are worse off at better colleges, but rather that the students who are admitted to top colleges are not the ones who would benefit the most from attending them.[7]

Thus the link between test score gaps and earnings inequality depends on: 1) the distribution of returns to college quality; and 2) how well exams predict these returns. I now turn to an empirical analysis that estimates heterogeneity in $_i$ from an admission exam reform.

2. Admission exam reform in Colombia

This section gives background on the Colombian college system and a reform of the national admission exam. I show that this reform dramatically reduced test score gaps between socioeconomic groups, but it also reduced the exam's predictive power for student outcomes.

2.1. **Institutional background and data.** I use three administrative datasets from the Colombian higher education system. First, I use records from a national standardized exam called the ICFES, which all Colombian students are required to take to apply to college. The Colombian exam is analogous to the U.S. SAT exam, but it is taken by nearly all high school graduates. I use individual administrative records from the testing agency that cover all 1998–2001 exam takers.[8] These provide each student's test scores and high school affiliation, as well as several measures of socioeconomic status (SES). I also use the testing agency's records on 2004–2011 takers of a field-specific college *exit* exam (described below).

Second, the Ministry of Education provided enrollment and graduation records for the near universe of colleges in the country. Colombia's college system consists of public and private institutions with varying selectivity and degree offerings. The Ministry's records cover

[7] Appendix B.2 shows that the usual measures of exam validity (equation (2)) are useful for maximizing earnings only if there is a strong correlation between returns to college quality, $_i$, and college outcomes, Y_i. Below I examine the relationship between returns and other college outcomes since it is difficult for testing agencies to observe $_i$ in practice.

[8] Individual-level exam data are not available before 1998. I exclude cohorts after 2001 because the sample of exam takers in my records changes, and the testing agency stopped collecting several SES variables.

students who enrolled in nearly all of these colleges in 1998–2012. The data include each student's institution, program of study, dates of entry and exit, and graduation outcome.

Finally, I use tax data from the Ministry of Social Protection, which contain monthly employment and earnings for any college enrollee working in the formal sector in 2008–2012.

I link the admission exam, college, and earnings records using individuals' names, birthdates, and ID numbers. The resulting dataset contains college outcomes for 1998–2001 exam takers and formal sector earnings for college enrollees measured in 2008–2012.[9]

2.2. **Variable definitions.** I define three measures of socioeconomic inequality from information collected at the time students took the admission exam:

(1) Gaps between the top and bottom quartiles of the family income distribution;[10]
(2) Gaps between students with college and primary (or less) educated mothers;
(3) Gaps between students who attended high and low ranked high schools.[11]

My main outcome variables are admission exam scores, college quality, college graduation, and earnings. I convert raw admission scores to percentiles within a student's exam cohort. I define college quality as a school's mean admission score across all pre-reform students and all subjects, and convert it to percentiles in the same manner. This is a fixed measure of quality that is common in work on college selectivity (Dale and Krueger, 2002; Hoxby, 2009), and it provides a natural link between admission exam design and college outcomes.[12] My graduation variable is an indicator for graduating from any college in the Ministry of Education records. For labor market outcomes, I use the log of an individual's average daily earnings measured 10–11 years after taking the admission exam.[13]

Table 1 shows summary statistics for these variables by SES group. I observe 1.6 million high school graduates who took the national admission exam in 1998–2001. Most of my analysis focuses on the 612,949 students who enrolled in college. Columns (C)–(F) show large gaps in all outcomes between SES groups. For example, students from the top income quartile scored 26 percentile points higher on the pre-reform math subject of the admission exam than bottom quartile students. Top quartile students also attended colleges ranked 20

[9] Appendix C.1 provides details on the coverage of each dataset and the merge.

[10] I define income quartiles relative to the population of college enrollees. In the data, family income is reported as a fraction of the monthly minimum wage, which changes over time due to inflation and policy. To obtain a stable measure, I compute predicted family income using individual characteristics and high school affiliation. See Appendix C.1 for details on this and my other SES measures.

[11] High school ranks are pre-reform measures from the ICFES testing agency. High school choice in Colombia is mainly determined by geography and tuition, and thus school rank is strongly related to family background.

[12] Appendix Table A1 and Appendix Figure A1 show that my measure of college quality is strongly related to both institutional levels and financial resources.

[13] Ten and eleven years after the test are the two experience levels at which I can measure earnings for all exam cohorts. I compute average daily earnings by dividing total annual earnings by the number of formal employment days in the year, demeaning by exam cohort and year, and averaging across the two years.

percentile points higher on average, were 14 percentage points more likely to graduate from college, and earned roughly 40 percent more one decade after the admission exam.

2.3. **The 2000 admission exam reform.** The Colombian national exam was first administered in 1968 with the aim of supporting college admissions. As the exam gained widespread coverage in the 1980s, the government began using its results to evaluate high schools. In the mid 1990s, policymakers concluded that the exam was poorly designed for the dual objectives of college admissions and high school accountability. The perception was that the exam rewarded test prep and memorization more than learned material. Critics argued that the exam was a poor measure of high school value added, and that scores were biased in favor of high SES students who had greater access to test prep services.

The testing agency overhauled the admission exam to address these concerns, altering the type of questions, the tested material, and the scoring system. The goal was to develop an exam that tested "competencies" rather than "content," and to align the exam with high school curricula. Appendix C.2 gives further details on changes to the exam structure and provides sample questions from the pre- and post-reform tests.

The new exam debuted in 2000 after more than five years of psychometric research. In preceding months, the exam overhaul was widely publicized in media outlets including *El Tiempo*, the leading Colombia newspaper. In its objective and publicity, the redesign of the exam was similar to the overhaul of the U.S. SAT exam in 2016.

2.4. **Reduction in SES test score gaps.** Figure 1 shows that the exam reform reduced the test score gap between students from the top and bottom income quartiles. The height of each bar is the difference in mean test percentiles between the top and bottom income quartiles. In pre-reform cohorts (1998–1999), the gap was approximately 27 percentile points in each of the six subjects of the exam. The reform had a dramatic impact on math and physics scores, with test score gaps falling by 50 percent in the post-reform cohorts (2000–2001). There were smaller reductions in language arts, biology, and chemistry test score gaps, and only minor effects in social science.

Table 2 shows that the reform also reduced test score gaps across other dimensions of SES. Column (A) displays mean test score gaps across the six exam subjects in pre-reform cohorts. These gaps ranged from 24 to 31 percentile points using family income, mother's education, and high school quality as measures of SES. Columns (B)–(G) show the change in test score gaps between the pre- and post-reform cohorts by exam subject. The reform had large effects in math and physics, with test score gaps falling by roughly 50 percent for each SES measure. The decline in test score gaps was more modest in other exam subjects, and only 1–3 percentile points in social science.[14]

2.5. **Reduction in exam validity.** Table 2 shows that the testing agency achieved its goal of reducing the relationship between socioeconomic status and exam scores. A second objective of the reform was to design a better measure of abilities that predict college success. Did the new exam reduce test score gaps while remaining a good predictor of college outcomes?

I address this question in two ways. First, I follow testing agencies' standard practice in measuring exam validity. The benchmark definition of validity is the correlation coefficient between test scores and college outcomes (Section 1.3). I use three college outcomes:

(1) Scores on a field-specific college *exit* exam that students take prior to graduation.[15]
(2) An indicator for graduating from any college in my administrative records.
(3) First-year GPA; I do not observe GPA in my administrative data, but I use a subsample of students in the 2000–2004 enrollment cohorts at a flagship university for whom I have transcript records (name withheld). See Appendix C.1 for details.

To reduce the influence of a student's college choice, I follow standard practice by using residuals from regressing both test scores and outcomes on college dummies (e.g., Kobrin et al., 2008). Thus the correlation coefficients reflect only within-college variation.

Second, I exploit heterogenous effects of the reform across exam subjects by using social science scores as a relatively fixed measure of ability. To measure exam validity, one would ideally observe students' performance on a separate test taken earlier in high school. Such data are often unavailable to testing agencies, nor do I have such a measure for this paper. As an alternative, I show how the reform affected the correlation of social science scores with other subject scores. Social science scores are not a perfect measure of ability because they reflect different skills than other subjects, and because the reform also altered social science questions. Nonetheless, the reform had a measurably smaller effect in social science (Table 2), and these scores partially measure a common component of ability.

Table 3 shows that the reform reduced the admission exam's predictive power for both college outcomes and social science scores. Column (A) shows how the (within-college) correlation of exit and admission exam scores changed from the pre- to the post-reform cohorts.[16] This correlation fell by about 0.15 points in math and physics—the subjects with the largest reductions in SES test score gaps. This is a 40 percent decrease in validity from

[14] The reform also reduced the gender gap in admission exam scores in several subjects, but I do not find that it affected gender gaps in college quality or other outcomes.

[15] The exit exam is now a national requirement for graduation, but it was optional during this time period. Roughly 40 percent of college enrollees in my sample took the exam, which reflects its voluntary nature and the fact that many students do not graduate. See MacLeod et al. (2017) for details on the exit exam.

[16] I normalize variables in Table 3 so that regression coefficients can be interpreted as correlation coefficients.

the mean pre-reform correlation of 0.39. The reform induced modest reductions in exit exam validity in language arts and biology, and had little effect in chemistry and social science.

Columns (B)–(D) show that the reform also reduced the exam's validity for graduation, GPA, and social science performance. The correlation between math/physics scores and graduation fell by 0.06 points, or roughly 50 percent of the pre-reform mean. Reform effects on graduation validity are mixed for other subjects. The GPA effects are underpowered because few students in my transcript data took the pre-reform exam. Nonetheless, the math and physics exams became weaker predictors of first-year GPA. Lastly, the reform drastically reduced the correlation of social science scores and other subject scores. In particular, correlations with math and physics scores fell by over 50 percent.

To explore heterogeneity in these effects, Panel A in Figure 2 plots the relationship between math and social sciences scores in each exam cohort. The horizontal axis is a student's percentile on the social science exam, and the vertical axis is the mean percentile on the math exam. Dark solid lines depict math scores for pre-reform cohorts (estimated by local linear regression). Lighter dashed lines plot math scores for post-reform cohorts. There is a dramatic increase in math scores at the bottom of the social science distribution, and a large reduction in math scores at the top. This suggests that the reform had effects across the distribution of ability, with the most impact on the highest and lowest ability students.

Since the reform may also have affected the distribution of social science scores, Panel B replicates the results in Panel A using the subset of students who took *both* the pre- and post-reform admission exams. The two lines in this panel are computed from the same sample, and in both cases the horizontal axis is the student's *pre-reform* social science score. The pattern of results is similar to that in Panel A. Performance on the math component of the new admission exam is much higher for students with low scores on the old social science exam, and much lower for students with high pre-reform social science scores.[17]

In sum, the Colombian exam reform reduced SES test score gaps, but at the cost of less predictive power for abilities that influence college outcomes. A potential explanation for these results is that the new exam questions were easier, and thus a larger proportion of the variance in performance came from randomness. The rest of this paper asks how the abilities measured by the pre- and post-reform exams relate to students' labor market outcomes.

3. Identification

This section presents my identification strategy for estimating the longer-run effects of the Colombian admission exam reform. I first discuss how public university admission methods created geographic variation in the stakes of the exam reform. I then define two treatment

[17] Appendix Figure A2 shows that the pattern of results in Figure 2 is similar if I use college *exit* exam scores—rather than social science *admission* scores—as a measure of ability.

variables that capture this variation. Lastly, I describe my differences-in-differences specification and show evidence consistent with the parallel trends identification assumption.

3.1. **College admission methods.** As in the U.S., college admissions in Colombia are decentralized. Students apply to individual colleges, and each institution controls its selection criteria. This decentralized system helps to identify the effects of the exam reform because only some colleges used the national exam for admissions.

Many selective colleges admit students solely based on national exam scores. These schools compute weighted averages of national exam subject scores and admit the highest ranking students up to a quota. As in many countries, Colombian college admissions are to school/major pairs, and the weights on each subject score usually vary across programs.

Other colleges do not consider national exam scores. Most commonly, these colleges require applicants to take the school's own exam, and use these scores as the sole admission criterion. Some colleges also consider other factors like high school grades or interviews.[18]

This paper focuses on admission methods at public universities, which were often the only selective colleges that low SES students could afford to attend. Colombia has roughly 50 public universities including a flagship school in most regions. Public schools are much less expensive than comparable private colleges and give tuition discounts to low SES students. Flagships are often considered the best college in the region and in some cases enroll over one-third of all local college students. Public universities also vary in admission methods. For example, the flagship school in Cali (Universidad del Valle) uses the national exam, while the flagship in Bogotá (Universidad Nacional de Colombia) administers its own entrance exam.

Colombia also has a network of elite private colleges, but their admissions were less important to low SES students. Financial aid markets were essentially non-existent at the time, and tuition was often at least five times higher at top private schools (Riehl et al., 2016). Thus low-income students rarely attended or even applied to selective private universities.[19]

Since most students attend college close to home, different admission systems at public universities created geographic variation in the stakes of the national exam reform. The reform was more consequential in areas where the local public university used only national exam scores. Since most public universities are large and highly selective, changes in their admission outcomes would also affect admissions at less selective colleges in the area. The reform mattered less in areas where the local public university used other admission criteria.

[18] One explanation for the variation in admission methods is that it is costly to administer a separate exam. Colleges that use the national exam tend to be in less-populated regions, and thus have lower budgets.

[19] As an illustration, the top private college (Universidad de Los Andes) admitted roughly half of its applicants during this time, while Universidad Nacional de Colombia's admission rate was near ten percent.

3.2. **Definition of treated areas.** To capture this variation, I define two geographic measures of reform "treatment" using the pre-reform admission criteria at public universities.[20]

First, I define *region-level* treatment based on public university admission methods in the administrative *departamento* where a student attended high school. I define a student as "treated" if she went to high school in a region where public universities used only national exam scores in admissions. "Control" students are from regions where public universities use their own entrance exams or other admission criteria. Colombia has 33 administrative regions, and in most cases this classification is unambiguous because the region contained only one public university and/or admission method. In the three regions with mixed admission criteria, I define treatment as the (enrollment-weighted) modal admission method.[21] For the sparsely-populated regions with no universities, I define treatment using the closest public university to the region's capital city.[22] Appendix Table A2 provides details on public university admission methods and this region-level treatment variable.

Second, I define *municipality-level* treatment based on the pre-reform propensity to enroll in national exam universities in each administrative *municipio*. This treatment variable equals the ratio of: 1) the number of 1998–1999 exam takers in a municipality who enrolled in a public university with national exam admissions; to 2) the number of 1998–1999 exam takers in that municipality who enrolled in any college. This is a continuous measure of treatment that reflects exposure to national exam universities in the municipality where a student attended high school. Colombia has roughly 1,000 administrative municipalities.

Figure 3 shows the region- and municipality-level treatment variables. Black dots in both panels represent public universities with national exam admissions, with dot sizes proportional to enrollment. White dots are public universities with other admission criteria. Panel A shows the binary region-level variable, with treated regions shaded dark red and control regions shaded light yellow.[23] Panel B shows municipality-level treatment, with darker colors reflecting higher pre-reform enrollment rates in national exam universities. Enrollment rates range from near zero in some municipalities to over 50 percent in urban areas near national exam universities. The correlation between the two treatment measures is 0.76. The mean of the municipality-level treatment variable is 36 percent in treated regions and five percent in control regions. Section 3.4 presents characteristics of treated and control areas.

[20] I obtained pre-reform admission methods from historical websites and regulations.

[21] Specifically, Bogotá is a control region because only 17 percent of public university students attended schools with national exam admissions. Caldas and Valle del Cauca are treated regions because 74 percent and 92 percent of students in public universities attended national exam universities.

[22] The main results are similar when I exclude small regions because they have few college students. Results are also similar when I exclude regions with mixed admission criteria (Bogotá, Caldas, and Valle del Cauca), although they are less precisely estimated. See Appendix Tables A14 and A15.

[23] The capital city of Bogotá is its own administrative region. Figure 3 does not display the Caribbean island region of San Andrés y Providencia. I define both of these areas as control regions.

3.3. **Differences-in-differences specification.** My empirical approach combines geographic variation in national exam stakes with the timing of exam reform. My benchmark specification is a standard differences-in-differences regression

$$y_{igt} = \ {}_g + \ {}_t + \theta(\text{Treated}_g \times \text{Post}_t) + u_{igt}, \tag{7}$$

where y_{igt} is an outcome for student i who attended high school in geographic area g, and took the national admission exam in cohort t. The regression includes geographic area dummies, ${}_g$, exam cohort dummies, ${}_t$, and the interaction between the treatment variable, Treatment_g, and a dummy for post-reform exam cohorts, Post_t. I run regressions with both region- and municipality-level treatment variables, and use region- and municipality-level dummies correspondingly. I cluster standard errors at the region level in all regressions.[24]

The coefficient of interest, θ, measures how outcomes changed with the reform in treated areas relative to control areas. In regressions with region-level treatment, θ reflects the average change in outcomes in treated regions relative to control regions. In municipality regressions, I normalize the treatment variable so that θ measures the effect of a one standard deviation increase in pre-reform enrollment at national exam universities. I discuss my outcomes variables and predictions for θ in Section 4.

3.4. **Identification assumptions and balance tests.** The main identification assumption is the usual differences-in-differences requirement of parallel trends. In this context, parallel trends means that college and labor market outcomes would have evolved similarly in treated and control areas in absence of exam reform. This section shows evidence of parallel trends in observable student traits, high school attainment rates, and macroeconomic conditions.

In Table 4, Panel A describes student characteristics in treated and control areas. Columns (A) and (B) display means for pre-reform exam takers in treated and control regions. Region-level treatment classifies 20 regions as treated and 13 as control. Control regions are more densely populated on average, with roughly 50 percent more exam takers per cohort. Control region students come from more socioeconomically advantaged backgrounds as measured by family income, parents' education, and attendance at top high schools.

While treated and control students have different characteristics, I find little evidence of differential trends in student traits across cohorts. Columns (C)–(F) display θ coefficients and standard errors from separate estimations of equation (7) using the dependent variable in the first column. Columns (C)–(D) show estimates with the region-level treatment variable, and columns (E)–(F) use municipality-level treatment. Point estimates are small and are statistically significant only for mother's education. I also present balance tests that combine all individual characteristics in my dataset (see Appendix Table C5) into a single index using

[24]Appendix Tables A12 and A13 show that the main results in this paper are robust to the wild t bootstrap procedure recommended by Cameron et al. (2008) for specifications with a relatively small number of clusters.

the predicted values from a log earnings regression. The estimated effects on this index are close to zero and are precisely estimated. These results suggest that changes in cohort composition were not systematically related to exposure to national exam universities.

Panel B presents evidence of parallel trends in economic conditions. Colombia experienced a severe recession in the late 1990s, and any heterogeneous effects of this recession could have caused divergent trends in college choices. To explore this possibility, Panel B present estimates of labor force participation and unemployment rates using labor market survey data.[25] Treated and control regions had relatively similar pre-reform employment conditions, and there is little evidence of differential trends during the recession.

Figure 4 corroborates this finding by plotting trends in high school graduation and unemployment rates from labor market surveys. Both panels display mean outcomes in treated and control regions, with a vertical line separating pre- and post-reform periods. Panel A shows high school graduation rates for cohorts defined by the year individuals turned 17—the most common age at which students take the college admission exam. High school graduation rates climbed rapidly in the 1990s and are higher in control regions, but they appear to have been on similar paths in the two areas. Panel B plots unemployment rates, which rose from ten percent in the mid 1990s to over 16 percent by the turn of the century. Yet treated and control areas had similar unemployment rates through the economic downturn. Note that I cannot rule out moderately-sized differences in employment trends due to volatility. Below I present graphical event studies to show that changes in my outcomes of interest are more plausibly related to exam reform than to concurrent macroeconomic conditions.

3.5. **Behavioral responses to exam reform.** This section examines potential behavioral responses to the exam reform, including effects on exam retaking, geographic mobility, and the probability of attending any college. Such responses are potential causal effects of the reform, but they would affect interpretation of the results in Section 4.

In Table 4, Panel C presents evidence on these potential behavioral responses. The reform had no significant effect on exam retaking or the probability of enrolling in any college. Although students in treated regions are more likely to retake the admission exam (columns (A)–(B)), changes in retaking rates are unrelated to the region- and municipality-level treatment variables (columns (C)–(F)). Overall college enrollment rates are similar in treated and control regions and did not diverge significantly with the reform.[26] A potential explanation

[25] Data are region-level estimates from the household survey *Encuesta Nacional de Hogares*. Individual and municipality-level survey data are not available prior to 2001.

[26] There is evidence that the concurrent recession reduced the overall number of tertiary enrollees (World Bank, 2003), but Panel C suggests that any enrollment declines were similar in treated and control areas. Appendix Table A5 also shows little evidence of differential changes in these outcomes by SES group.

for this result is that Colombia has many non-selective colleges where students can enroll if they are not admitted to top colleges.[27]

There is some evidence that the reform affected student mobility across regions. There are no significant changes in the mean distance between a student's high school and college, but I find a 1–2 percentage point increase in the relative probability that treated students stayed in region for college. I also a find a marginally significant increase in the fraction of treated region college enrollees who attended public universities. These results suggest that there was a relative increase in student body size at public universities in treated areas

The reform may have induced some students to go to college closer to home, but the magnitude of this effect is small relative to normal fluctuations in cohort size. The largest estimate in Table 4 equates to 90 extra students per cohort at each public university. During this period, the mean cohort size at these schools was about 2,000 students, and the within-school standard deviation was 400 students. Further, most Colombian students do not enter college right after high school, so any class size effects were spread out over multiple cohorts. This suggests that the results below are unlikely to be driven by overcrowding effects.

4. Exam reform effects on college quality and earnings

This section shows how the Colombian admission exam reform affected college and labor market outcomes. I find that the reform reduced SES gaps in college quality, but this reallocation decreased graduation rates and earnings for *both* high and low SES students.

4.1. **Effects on the SES college quality gap.** To explore the exam reform's effects on college choices, I modify the differences-in-differences regression (7) to estimate changes in the *gap* in college quality between high and low SES students. For this, I use college quality, Q_c, as the dependent variable, and I interact all right-hand side variables with a dummy for high SES students, X_i. This yields the triple differences regression:[28]

$$Q_c = \ _{gt} + \Big(\ _g + \ _t + \theta^q(\text{Treated}_g \times \text{Post}_t)\Big)X_i + u_{icgt}.\psi \tag{8}$$

The coefficient of interest, θ^q, measures the change in the SES college quality gap in treated areas relative to control areas. As above, Q_c is a fixed measure that equals a college's percentile rank based on mean pre-reform admission scores, so θ^q is in percentile units. Since

[27] A similar finding emerges in work on other large-scale policies that primarily affect admission to selective colleges. For example, Hinrichs (2012) and Daugherty et al. (2014) show that affirmative action bans and percent plan admission rules have little effect on the extensive margin of college enrollment.

[28] This specification is equivalent a two-step estimation procedure (e.g., Card and Krueger, 1992): 1) estimate the college quality gap in area g and cohort t; and 2) use this gap as the dependent variable in the differences in differences regression (7). Specification (8) is the result of plugging the first step into the second. See Appendix C.3 for a derivation of this specification from the model equations in Section 1.

the reform lowered test score gaps (Section 2), the prediction is that it should also lower the SES college quality gap in areas where the exam is important for admission ($\theta^q < 0$).

Table 5 presents estimates of θ^q from equation (8). Column (A) uses the region-level treatment variable with region dummies ($_g$ and $_{gt}$). Column (B) uses municipality-level treatment with municipality dummies. Each coefficient is from a separate regression that defines high SES, X_i, by family income, mother's education, or high school rank.

Columns (A)–(B) show that the reform reduced SES college quality gaps in treated areas relative to control areas. The first row of column (A) shows that the relative college quality gap by family income fell 2.4 percentile points in treated regions from a base of roughly 20 percentile points (Table 1). The magnitude is similar using municipality-level treatment; a one standard deviation increase in pre-reform exposure to national exam universities is associated with a 1.6 percentile point decrease in the college quality gap by family income.[29] The effects are smaller in magnitude using mother's education and high school rank as measures of inequality, though in all cases the reform reduced college quality gaps.

Figure 5 shows the timing of the effects in Table 5. Panel A plots the mean college quality gap between top and bottom income quartile students in treated regions (dashed line) and control regions (solid line). The left and right vertical axes are shifted so that the treated and control region gaps match in 1999—the last cohort before the exam overhaul. Panel B plots coefficients θ_t^q from event-study versions of equation (8) using the municipality-level treatment variable. These coefficients calculate separate college quality effects for each exam cohort t, omitting the 1999 cohort. Vertical dashed lines are 95% confidence intervals.

Both panels show that the college quality gap in treated areas fell in the first cohort after the reform. Panel A shows no evidence of differential pre-trends in the two pre-reform cohorts (1998–1999). In the first cohort after the reform, the college quality gap fell sharply in treated regions. By the 2001 cohort, the college quality gap declined by more than two percentile points in treated regions relative to control regions. A similar pattern arises in Panel B using event-study coefficients from the municipality-level treatment variable.[30]

4.2. **Robustness tests for college quality effects.** To address potential concerns about differences between treated and control areas, I add controls to the benchmark regression so that identification is restricted to municipalities with similar populations. Control students tended to live in more populated areas (Table 4), so their colleges choices may have been differentially affected by labor market conditions in the concurrent recession. To address this, I use hierarchical clustering to divide municipalities into ten groups based on the number of exam takers. I also create a separate group for the three largest cities: Bogotá, Medellín,

[29] One standard deviation is 18 percentage points in pre-reform national exam university enrollment.

[30] Appendix Figure A3 presents graphs similar to Figure 5 for college quality gaps defined by mother's education and high school rank. Results are similar, but there is a pre-trend in mother's education gaps.

and Cali.[31] I add into specification (8) dummies for these eleven population groups fully interacted with cohort and SES dummies (e.g., $_{Ktx}$ dummies, where $K\psi$ defines population groups and $x\psi$ defines SES groups). With these controls, identification comes from municipalities with similar populations but different treatment values. For example, the reform's effects in Cali—the largest treated region city—are identified only by comparison with Bogotá and Medellín—the largest control region cities.

Columns (C)–(D) in Table 5 show that the college quality effects persist with the addition of population controls. The decline in the family income gap is *larger* in magnitude when one compares only municipalities with similar populations. The estimated effects on college quality gaps by mother's education and high school rank remain negative, although the magnitudes fall and standard errors increase.

Columns (E)–(F) add extra controls for proximity to public universities. Prior work shows that distance to college is a strong predictor of college attendance (e.g., Card, 1993). Ex-ante differences in proximity to public universities could lead to divergent trends in treated and control areas given Colombia's rising educational attainment. To address this, I divide municipalities into ten groups based on distance to a public university using the same clustering method as above. I then add dummies for full interactions between proximity groups, population groups, exam cohorts, and SES groups (e.g., $_{JKtx}$ dummies, where $J\psi$ defines proximity groups) into specification (8). With these controls, college quality effects are identified within groups of 14 municipalities on average, where all municipalities in a given group have similar populations and proximities to public universities.

The results in columns (E)–(F) are similar to those in columns (C)–(D), suggesting differences in university access are not driving the main results. The proximity controls have explanatory power for students' college quality, as standard errors increase in all six regressions. But coefficients are similar to those in the previous specification.

Appendix Tables A3–A4 show that the reform had little impact on students' major choices. Colombian students apply to college-major pairs, so it is possible that the exam reform affected the distribution of majors. Appendix Table A3 is similar to Table 5, but it uses major selectivity—rather than college quality—as the dependent variable. I find no effects on SES gaps in major selectivity, suggesting that the reform mainly affected the schools students attended. Appendix Table A4 also shows little evidence of changes in major choice by SES using indicators for major groups as dependent variables.

4.3. **Effects on graduation and earnings.** This section shows how the reallocation of college quality affected students' graduation and labor market outcomes. Table 6 presents

[31] Appendix Table A6 shows the eleven population groups. I use Ward's method for hierarchical clustering, which minimizes the total within-cluster variance given the choice of K clusters. I do not use simple deciles to define groups because the distribution of municipality populations is highly skewed.

estimates of θ from the standard differences-in-differences regression (7). I use three different dependent variables: an indicator for college persistence, defined as still being enrolled one year after starting college (columns (A)–(B)); an indicator for college graduation (columns (C)–(D)); and log daily earnings measured 10–11 years after taking the admission exam (columns (E)–(F)). The odd columns use the region-level treatment variable, and the even columns use municipality-level treatment. The first row shows estimates including all students, which measures the overall effect of the reallocation of college quality. The remaining rows show separate estimates for different groups of high and low SES students.

The reform lowered overall graduation rates and post-college earnings in treated areas relative to control areas. In the first row, the region-level estimates show that graduation rates declined by 1.5 percentage points in treated regions relative to control regions, and mean earnings fell by 1.3 percent. The magnitude of the reform effects are similar using municipality-level treatment; a one standard deviation increase in exposure to national exam universities is associated with a 0.8 percentage point decline in graduation rates, and a 0.7 percent decrease in mean earnings. Columns (A)–(B) show that graduation effects came primarily from students dropping out by the end of their first year.[32]

Figure 6 corroborates these results by showing that mean earnings decreased in the first cohort after the exam reform. Panel A plots average log daily earnings in treated regions (dashed line) and control regions (solid line), with axes shifted so that earnings match in 1999. Panel B plots θ_t coefficients from an event-study version of equation (7) using municipality-level treatment. There is little evidence of a pre-trend in earnings using either region- or municipality-level treatment. In both cases, there is a decline in market-wide earnings in the first post-reform cohort. Appendix Figure A4 shows similar patterns for college persistence and graduation rates. The sharp changes in treated area outcomes suggest that these effects are related to the exam overhaul rather than to concurrent macroeconomic events.

In Table 6, the second and third rows show that the decline in graduation rates and earnings came from *both* high- and low-income students. Across all outcomes and treatment variables, the reform effects are negative for students in both the top and bottom income quartiles. Further, the magnitudes of these effects are relatively similar between the two income groups. For example, relative earnings fell by 1.8 percent for high-income students in treated regions, and by 1.3 percent for low-income students in treated regions.

The remaining rows in Table 6 show similar results for SES groups defined by mother's education and high school rank. In all cases the estimated effects are negative for both high and low SES students, with similar magnitudes between the two groups.

[32] The magnitudes of the persistence, graduation, and earnings effects are consistent with a college earnings premium of roughly one log point, which is close to the cross-sectional premium in Colombia survey data.

4.4. **Robustness tests for graduation and earnings effects.** Appendix Tables A7–A9 show that the graduation and earnings results in Table 6 are robust to the inclusion of population and college access controls. These tables present results analogous to columns (C)–(F) in Table 5; I add into equation (7) dummies for municipality population and proximity groups interacted with exam cohort dummies. Including these controls increases standard errors, but magnitudes are similar and the estimates remain statistically significant in most cases. The graduation and earnings effects by SES group are negative in all regressions.

The labor market results in Table 6 are also robust to different earning measures. My benchmark results define earnings at a fixed time since the admission exam (10–11 years later). An alternative approach is to measure earnings in the same years but at different levels of post-exam experience. Appendix Table A10 shows that I find similar results for earnings measured in calendar years 2011–2012.[33] I also find similar results using *annual* rather than *daily* earnings, although the estimates are less precise. This suggests that the reform did not have a significant effect on the number of days that individuals worked.

Appendix Table A11 shows that the reform may have cause a small decrease in the rate of formal employment. I only observe earnings for individuals who work at firms that are registered with the Ministry of Social Protection (56 percent of my sample).[34] Appendix Table A11 shows estimates from equation (7) in which the dependent variable is an indicator for appearing in the earnings data. Estimates are insignificant using region-level treatment, but a one standard deviation increase in municipality-level treatment is associated with a one percentage point reduction in having any earnings. These results are consistent with an increased rate of informal or non-employment as a result of lower educational credentials. This is another potential effect of the reform, but it is unlikely to change the sign of the earnings estimates in Table 6 because wages are typically much lower in the informal sector.

4.5. **Interpretation of results.** In interpreting the graduation and earnings effects, it is important to note that the reform was likely more consequential from the perspective of students than from the perspective of colleges. The magnitudes of the college quality effects in Table 5 suggest that the reform induced roughly 1,000 low SES students per cohort to move up a college quality tier (see Appendix C.4). These effects were spread out across more than 100 treated region colleges and over multiple enrollment cohorts. Thus it is unlikely that colleges experienced a perceptible change in student composition in the initial years after the reform. To formalize this point, Appendix Table A16 shows that the reform had no effect on

[33] Measuring earnings in the same years for all exam cohorts also helps to alleviate potential concerns about differential labor market conditions in treated and control areas.

[34] Individuals who do not appear in my earnings records can be either informally employed, unemployed, or out of the labor force. In 2008–2012 survey data (*Gran Encuesta Integrada de Hogares*), about 30 percent of workers with any college education were employed in jobs with no contract.

the correlation between college quality, Q_c, and observable cohort peer characteristics (e.g., mean SES or mean social science score). A caveat is that I cannot rule out larger effects on the distributions of unobserved ability across colleges.

The small magnitudes of the college quality results argue against peer effects or institutional responses as a potential mechanism for the results in Table 6. These results are more likely to be explained by individuals having different returns at different colleges.

The negative outcomes for high SES students are consistent with reductions in their college quality. As in the U.S., less selective colleges in Colombia have fewer resources and lower graduation rates (see Appendix Table A1 and Figure A1). Thus effects for high SES students are consistent with work showing that lower resources leads to lower graduation rates (Bound et al., 2010; Deming and Walters, 2017). Further, they suggest that decreased educational attainment harmed these students in the labor market.

The negative effects for low SES students in treated areas are more striking because these individuals experienced *increases* in college quality on average. The reduction in low SES graduation rates and earnings is consistent with "mismatch" (Arcidiacono and Lovenheim, 2016) in that some of these students may have been better off at lower ranked colleges. I explore this point more formally in the next section.

5. Returns to college quality

This section estimates returns to college quality from the Colombian exam reform. I first use the above results to develop an instrumental variables (IV) specification that estimates average returns by SES group. Next, I use new machine learning techniques to estimate heterogeneity in returns within SES groups. Finally, I show that the reform reduced the correlation of admission scores and returns, and I relate this result to Proposition 1.

5.1. **Instrumental variables specification.** Section 1 highlighted the importance of students' returns to college quality, defined as $\ _i$ coefficients from the earnings equation (5):

$$w_{ic} = \ _i + \ _i Q_c + \epsilon_{ic}.\psi$$

A challenge in estimating $\ _i$ is that a student's choice of college quality, Q_c, is likely to be related to her potential earnings. To address this endogeneity, I use only geographic and cohort variation to estimate returns from the Colombian reform, as in the above differences-in-differences design.

My specification for returns to college quality is the two-stage least squares (2SLS) system:

$$Q_c = \ _{gz} + \ _{tz} + \theta_z(\text{Treated}_g \times \text{Post}_t) + u_{icgtz} \tag{9}$$

$$w_{icgtz} = \pi_{gz} + \pi_{tz} + \ _z \hat{Q}_c + \epsilon_{icgtz}.\psi \tag{10}$$

The first stage regression (9) is similar to equation (7) with college quality, Q_c, as the dependent variable. It includes dummies for geographic areas g and exam cohorts t, and the instrument is the interaction between the treatment variable and an indicator for post-reform cohorts, $\text{Treated}_g \times \text{Post}_t$. In the second stage regression (10), the dependent variable is log daily earnings, w_i, and the endogenous college quality variable is replaced with first stage predicted values. The return to college quality is the coefficient on $\hat{Q}_c$ in equation (10).

The 2SLS system differs from equation (7) in that all coefficients are interacted with dummies for groups defined by covariate values $Z_i = z$. Equations (9)–(10) thus estimate first stage effects, θ_z, and returns to college quality, $_z$, separately for each covariate group z. I use different covariate groups to estimate returns that relate to the parameters from equation (6) in the model. One model parameter is the average return to college quality for SES group x, $\bar{}_x$. Defining covariate groups as income quartiles, for example, provides an estimate of the average return for students in income quartile z. Another model parameter is the within-SES covariance between returns and test scores, $\text{Cov}(_i, T_i | X_i = x)$. Using additional covariates (e.g., gender and age) allows me to estimate heterogeneity in returns within SES groups, which I then correlate with exam scores.

5.2. **IV assumptions and interpretation.** Under the IV assumptions in Imbens and Angrist (1994), the parameter $_z$ can be interpreted as a local average treatment effect (LATE). Specifically, $_z$ is the return to college quality for reform "compliers" in group z—students whose college quality was altered by the exam reform. Complier average returns may differ from population average returns, but Proposition 1 carries through if model parameters are interpreted as defining complier outcomes. Thus the $_z$ estimates are informative for policies that produce similar complier populations as the Colombian reform.[35]

One assumption that is necessary for a LATE interpretation is instrument relevance, which in this case means that the reform altered the college quality of group z. In practice this means estimates of $_z$ may be biased for covariate groups whose college quality did not change substantially with the reform. Below I show how machine learning methods help address weak instrument issues that arise from an arbitrary definition of covariate groups.

A LATE interpretation also requires exclusion and monotonicity assumptions, which are unlikely to be strictly satisfied in this context. The exclusion restriction states that the exam reform affected students' earnings only through changes in their college quality as defined by Q_c. This assumption may not hold because Q_c is a one-dimensional measure of quality, and returns to college quality can vary on multiple dimensions. Monotonicity implies that

[35] In principle one could exploit the exam reform to compute other policy relevant parameters in a marginal treatment effects framework (e.g., Heckman and Vytlacil, 2005; Mogstad et al., 2019). In practice this analysis is complicated by the fact that treatment is defined by a continuous variable, Q_c, and by the instrument relevance issues discussed below.

the reform shifted college quality in the same direction for all students in group z. This may be violated given substantial uncertainty and preference heterogeneity for students choosing between a large number of college options (Carneiro et al., 2003; Kirkebøen et al., 2016).

Despite the potential violation of these assumptions, estimates of $_z$ are useful summary statistics for how the benefits of attending a better school vary across different types of students. Since testing agencies produce a one-dimensional ranking of students, it is useful to know how this ranking maps into earnings using a test-score measure of college quality.

5.3. **Average returns to college quality by SES.** Table 7 presents estimates from the 2SLS regressions (9) and (10) using the region-level treatment variable.[36] To ensure that the first and second stage regressions are estimated on the same population, the sample only includes students with formal sector earnings (column (A)).

Panel A shows estimates of the mean return to college quality for each SES group. In this panel, I define covariate groups $z\psi$ using only the SES variable listed in the first column. The estimates are equivalent to those from separate estimations of (9)–(10) for each SES group. The resulting coefficients are estimates of the average college quality effect, θ_z, and the average return to college quality, $_z$, for each SES group.

Columns (B)–(C) show first stage effects, θ_z, and standard errors from equation (9). I divide the dependent variable, Q_c, by ten so that one unit corresponds to ten percentile points in the distribution of college quality. Consistent with Table 5, the reform induced roughly a one percentile point increase in college quality for low SES students, and a one percentile point decrease for high SES students. As above, the reform had the biggest impact on the college quality gap by family income.

Columns (E)–(F) show returns to college quality, $_z$, and standard errors from the second stage regression (10). These coefficients are equal to Wald estimands that divide the reduced form earnings effects (column (E) in Table 6) by the first stage effects (column (B) in Table 7). Since the reform reduced both college quality and earnings for high SES students, their average return to college quality is positive. Low SES students also experienced decreases in earnings, but on average their college quality *increased* with the reform. Thus the mean return is negative for low SES students. These returns are statistically significant for the family income groups. A ten percentile increase in Q_c caused a nine percent increase in earnings for top-quartile students, and a nine percent decrease for bottom-quartile students.

The returns reported in Panel A should be interpreted with caution because first stage $F\psi$ statistics are low. Column (D) shows that $F\psi$ statistics from tests of $\theta_z = 0$ are below the rule of thumb value of ten for all SES groups except the top income quartile. This suggests a

[36] Appendix Table A17 shows estimates using municipality-level treatment.

failure of instrument relevance, which arises from an arbitrary definition of covariate groups z. I address this issue in the next section.

5.4. **Heterogeneity in returns to college quality.** This section estimates across- and within-SES variation in returns to college quality. I follow recent work that develops methods for estimating heterogeneity in causal effects (Wager and Athey, 2017; Athey et al., 2017). I adapt these methods to my empirical setting and to address the weak instrument issue noted above.

Specifically, I use Athey et al. (2017)'s random forest algorithms for estimating heterogeneity in IV moment conditions. In my context, the authors' algorithm provides a data-driven method for selecting covariate groups z in the 2SLS regressions (9)–(10). Given a vector of covariates Z_i, this algorithm searches for regression tree nodes $Z_i = z$ that maximize heterogeneity in the parameters θ_z and $_z$, while subsetting the data in a way that provides unbiased estimates of heterogeneity.[37] Appendix C.5 describes my implementation of the Athey et al. (2017) estimation procedure in detail. I briefly summarize it here.

I first define a vector of covariates Z_i that includes both SES measures and other individual characteristics, including gender, age, family structure, and high school type.[38] Since my identification relies on differences-in-differences variation, I regress each of my three main variables—log earnings (w_i), college quality (Q_c), and the instrument ($\text{Treated}_g \times \text{Post}_t$)—on region dummies and cohort dummies. I use residuals from these regressions to define the standard three-variable IV moment conditions. I estimate Athey et al.'s algorithm on a random 50 percent "training" subsample of my data. The algorithm uses covariates Z_i to define groups z and estimate the parameters θ_z and $_z$, with the goal of maximizing heterogeneity in these coefficients. Lastly, I predict values of these parameters into the other half of the data—the "validation" sample—which I use for the results below.[39]

Panel B of Table 7 summarizes the random forest estimates of θ_z and $_z$. This panel is similar to Panel A, but reported values are summary statistics across many parameter estimates. I compute these statistics separately for the SES groups listed in the first column.

[37] The degree of heterogeneity can be overstated if one searches over covariates for significant differences in treatment effects. Athey et al. (2017) call this algorithm an *instrumental forest* because it uses random forests to predict IV moment parameters. Their data subsetting procedure—*honest estimation*—uses one sample to draw regression trees, and another to estimate values at each node.

[38] Specifically, Z_i includes gender, age, birth order, number of siblings, number of siblings in college, family income, mother's and father's education, father's occupation, number of employed family members, high school rank, high school academic level, and high school ownership status.

[39] The main implementation choice is the minimum allowable size of covariate groups z. Large groups limit the ability to estimate heterogeneity in $_z$, while small groups reduce precision. I choose node size parameters that maximize the *sum* of F statistics from tests that each $_z$ is equal to the mean of the $_z$ estimates. This balances these considerations by rewarding both the number and the precision of $_z$ estimates.

Column (B) shows the mean of the θ_z estimates for each SES group, and column (C) shows the average standard error. The mean first stage effects in Panel B are similar to the single group estimates in Panel A, with average increases in college quality for low SES students, and decreases for high SES students. The mean standard error is roughly the same magnitude as the average coefficient, which shows that many first stage estimates are statistically insignificant. This is consistent with the low F-statistics in Panel A; the exam reform is not a strong instrument for college quality in many covariate groups.

To address this issue, I only present returns to college quality, $\ _z$, for covariate groups z with a strong first stage. Column (D) shows the number of students in covariate groups with first stage F-statistics greater than ten. This keeps about three percent of the sample—groups for whom the reform had the strongest effect on college quality.[40]

Column (E) presents the mean $\ _z$ estimate in this subsample of "reform compliers," and column (F) shows the average standard error. Mean returns are similar to the single group estimates in Panel A. For high SES students, a ten percentile point increase in college quality caused a 12.5 percent increase in earnings on average. For low SES students, a ten percentile point increase in college quality *reduced* earnings by 7–10 percent. This does not imply that the mean return to college quality is negative in the full population of low SES students. Rather, the low SES students most affected by the Colombian reform had negative returns.

Columns (G)–(H) examine heterogeneity in returns to college quality *within* SES groups. There are roughly 200 coefficients per SES group (column (G))—a result of the large vector of covariates Z_i used in estimation. Column (H) shows standard deviations of the $\ _z$ estimates. There is substantial heterogeneity in returns, with an overall standard deviation of 0.11 log points.[41] Within-SES standard deviations range from 0.04 to 0.09 log points. Variation in returns is smaller for high SES groups, suggesting that most high SES students affected by the reform had positive returns to college quality. For low SES students, standard deviations are close in magnitude to the mean return. Thus many low SES students who were affected by the reform had negative returns to college quality, but there is significant variation. I conclude by showing how this heterogeneity relates to pre- and post-reform exam scores.

5.5. **Predictive power of admission scores for returns to college quality.** This section shows that the Colombian reform reduced the admission exam's predictive power for students' returns to college quality, and discusses the implications for Proposition 1.

Figure 7 shows the relationship between math admission exam scores and returns to college quality in the pre- and post-reform cohorts. The x-axis in each panel depicts random forest

[40] The remaining sample is small because of the desire to explore heterogeneity in $\ _z$. For example, top income quartile students have a strong first stage as a single group (Table 7, Panel A), but many subgroups have weak first stages. Appendix Table C7 shows how each covariate relates to the θ_z and $\ _z$ estimates.

[41] I strongly reject the hypotheses that all $\ _z = 0$ ($F = 2.41$) and all $\ _z = E[\ _z]$ ($F = 2.16$).

estimates of returns, $_z$, for covariate groups z with first stage F-statistics above ten. The y-axis depicts the mean math percentile in each covariate group. Panel A uses pre-reform exam cohorts (1998–1999), and Panel B uses post-reform cohorts (2000–2001). The solid line is the linear relationship between these two variables for all students, with covariate cells represented by hollow circles. The red dashed line is the same relationship using only students from the bottom income quartile, with cells represented by triangles.

Figure 7 shows that the reform reduced the math exam's predictive power for students' returns to college quality. There is a strong positive relationship between pre-reform scores and returns for all students in the sample. This relationship declines with the reform, as evidenced by the flattening of the solid line. There is a similar pattern using only students in the bottom income quartile. The dashed line shows a smaller but positive relationship between low-income math scores and returns to college quality in pre-reform cohorts. This correlation falls to zero in post-reform cohorts.

To formalize these results, Table 8 presents estimates from regressions of returns to college quality on admission scores. Coefficients are from separate regressions for each exam subject. Column (A) shows the correlation between scores and returns in pre-reform cohorts. Column (B) shows the change in this correlation from pre- to post-reform cohorts. Columns (C)–(D) are analogous, but regressions include only students in the bottom income quartile. Variables are normalized so that coefficients can be interpreted as bivariate correlations.[42]

Table 8 provides further evidence that the reform reduced the exam's predictive power for returns to college quality. The correlation between pre-reform scores and returns is about 0.4 in each subject. This correlation falls by nearly 50 percent in the post-reform math and physics exams—the subjects with the largest decline in test score gaps and validity (Section 2). Results for the bottom income quartile are noisily estimated due to the small sample size. Nonetheless, the within-SES correlation of returns and pre-reform scores is positive in all subjects, and the exam's predictive power declines with the reform in math and physics.

These results connect the earnings effects of the reform to the framework in Section 1. Proposition 1 showed how reducing SES test score gaps affects earnings under different distributions of returns to college quality, $_i$. The results in Table 7 reject homogeneous returns (part (A)), and are instead consistent with a complementarity between SES and returns to college quality (part (B)). The strong relationship between SES and $_z$ for the most affected students explains the overall negative earnings effect of the reform.

[42] The notes to Table 8 describe how I adjust standard errors because $_z$ is estimated. This adjustment assumes no correlation between standard errors and $_z$ residuals, and I do not adjust for potential intra-cluster correlation of estimation errors. These assumptions do not affect the qualitative takeaways from Table 8 given the large t-statistics in columns (A)–(B) and the small t-statistics in columns (C)–(D).

The above results also suggest two types of "mismatch" for low SES students, consistent with part (C) of Proposition 1. First, the mean return to college quality was negative for the low SES students most affected by the reform (Table 7, column (E)). This is consistent with the usual definition of mismatch (Arcidiacono and Lovenheim, 2016)—some of these students may have been better off at lower ranked colleges. Second, the post-reform exam was a worse predictor of returns to college quality (Table 8), making it less likely that students with higher returns attended top colleges.

6. Conclusion

A large literature asks how attending a more selective college affects an individual's career prospects (e.g., Dale and Krueger, 2002). This work helps to explain why families expend a great deal of energy on college admissions (Ramey and Ramey, 2010), as it explores the impacts of college choice from the perspective of individuals.

This paper explored college choice from a market perspective. It asked how the matching of students to colleges via an admission test affects the distribution of earnings. Using a simple framework and a natural experiment, it showed that the earnings implications of an admission test depend on the exam's predictive power for students' returns to college quality. Exams with low predictive power for these returns can lead to mismatch in the assignment of students to colleges. This was the case in Colombia, where an admission exam overhaul reduced test score gaps but also the information content of the exam. The result was lower post-college earnings for both high and low SES students.

These results suggest that one should be cautious in extrapolating estimates from marginally admitted students to large-scale admission policies. Returns may differ for students whom a college deems to be qualified and for those who would not be admitted under status quo policies. This can explain why one might find large positive returns for disadvantaged students who are marginally admitted to top colleges (Hoekstra, 2009), and negative returns for students admitted via affirmative action (Arcidiacono et al., 2016).

Another takeaway is that increasing "fairness" in admission tests does not necessarily reduce inequality in the labor market. Reforms that lower test score gaps may only help disadvantaged students if there are other ways of identifying which students would excel at top colleges. In the U.S., a growing number of colleges have adopted test-optional admission policies. The results in this paper suggest it is important for such schools to use other admission criteria that replace the information content of exams.

A caveat to these conclusions is that non-pecuniary benefits of diversity can also justify efforts to reduce the role of socioeconomic status in admissions. Further, colleges may be able to adapt support services or teaching practices over time to improve student outcomes.

References

Abdulkadiroğlu, A., P. Pathak, and A. Roth (2005). The New York City high school match. *American Economic Review 95*(2), 364–367.

Andrews, R. J., J. Li, and M. F. Lovenheim (2016). Quantile treatment effects of college quality on earnings. *Journal of Human Resources 51*(1), 200–238.

Arcidiacono, P., E. Aucejo, P. Coate, and V. J. Hotz (2014). Affirmative action and university fit: Evidence from Proposition 209. *IZA Journal of Labor Economics 3*(1), 1.

Arcidiacono, P., E. M. Aucejo, H. Fang, and K. I. Spenner (2011). Does affirmative action lead to mismatch? a new test and evidence. *Quantitative Economics 2*(3), 303–333.

Arcidiacono, P., E. M. Aucejo, and V. J. Hotz (2016). University differences in the graduation of minorities in STEM fields: Evidence from California. *American Economic Review 106*(3), 525–562.

Arcidiacono, P. and M. Lovenheim (2016). Affirmative action and the quality-fit tradeoff. *Journal of Economic Literature 54*(1), 3–51.

Athey, S., J. Tibshirani, and S. Wager (2017). Generalized random forests. Forthcoming, Annals of Statistics.

Backes, B. (2012). Do affirmative action bans lower minority college enrollment and attainment? evidence from statewide bans. *Journal of Human Resources 47*(2), 435–455.

Bagde, S., D. Epple, and L. Taylor (2016). Does affirmative action work? caste, gender, college quality, and academic success in India. *American Economic Review 106*(6), 1495–1521.

Bertrand, M., R. Hanna, and S. Mullainathan (2010). Affirmative action in education: Evidence from engineering college admissions in India. *Journal of Public Economics 94*(1), 16–29.

Bettinger, E. P., B. J. Evans, and D. G. Pope (2013). Improving college performance and retention the easy way: Unpacking the ACT exam. *American Economic Journal: Economic Policy 5*(2), 26–52.

Bound, J., M. F. Lovenheim, and S. Turner (2010). Why have college completion rates declined? an analysis of changing student preparation and collegiate resources. *American Economic Journal: Applied Economics 2*(3), 129–57.

Bulman, G. (2015). The effect of access to college assessments on enrollment and attainment. *American Economic Journal: Applied Economics 7*(4), 1–36.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics 90*(3), 414–427.

Canaan, S. and P. Mouganie (2015). Returns to education quality for low-skilled students: Evidence from a discontinuity. Working Paper. Available at SSRN 2518067.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. National Bureau of Economic Research.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*(5), 1127–1160.

Card, D. and A. B. Krueger (1992). Does school quality matter? returns to education and the characteristics of public schools in the United States. *Journal of Political Economy 100*(1), 1–40.

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college. *International Economic Review 44*(2), 361–422.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating marginal returns to education. *The American Economic Review 101*(6), 2754–2781.

Chetty, R., J. N. Friedman, E. Saez, N. Turner, and D. Yagan (2017). Mobility report cards: The role of colleges in intergenerational mobility. Technical report, National Bureau of Economic Research.

Cortes, K. E. (2010). Do bans on affirmative action hurt minority students? evidence from the Texas Top 10% Plan. *Economics of Education Review 29*(6), 1110–1124.

Cullen, J. B., M. C. Long, and R. Reback (2013). Jockeying for position: Strategic high school choice under Texas' Top Ten Percent plan. *Journal of Public Economics 97*, 32–48.

Dale, S. B. and A. B. Krueger (2002). Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics 117*(4), 1491–1527.

Daugherty, L., P. Martorell, and I. McFarlin (2014). Percent plans, automatic admissions, and college outcomes. *IZA Journal of Labor Economics 3*(1), 1.

Deming, D. J. and C. R. Walters (2017). The impact of price caps and spending cuts on US postsecondary attainment. National Bureau of Economic Research.

Dillon, E. W. and J. A. Smith (2017). Determinants of the match between student ability and college quality. *Journal of Labor Economics 35*(1), 45–66.

Dillon, E. W. and J. A. Smith (2018). The consequences of academic match between students and colleges. NBER Working Paper No. 25069.

Durlauf, S. N. (2008). Affirmative action, meritocracy, and efficiency. *Politics, Philosophy & Economics 7*(2), 131–158.

Goodman, J., O. Gurantz, and J. Smith (2018). SAT retaking and inequality in college enrollment. Working Paper.

Goodman, S. (2016). Learning from the test: Raising selective college enrollment by providing information. *Review of Economics and Statistics 98*(4), 671–684.

Hastings, J. S., C. A. Neilson, and S. D. Zimmerman (2013). Are some degrees worth more than others? Evidence from college admission cutoffs in Chile. National Bureau of Economic Research Working Paper 19241.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73*(3), 669–738.

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

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

Hoekstra, M. (2009). The effect of attending the flagship state university on earnings: A discontinuity-based approach. *The Review of Economics and Statistics 91*(4), 717–724.

Hoxby, C. and S. Turner (2013). Expanding college opportunities for high-achieving, low income students. Stanford Institute for Economic Policy Research Discussion Paper.

Hoxby, C. M. (2009). The changing selectivity of American colleges. *The Journal of Economic Perspectives 23*(4), 95–118.

Hoxby, C. M. (2018). The productivity of U.S. postsecondary institutions. In C. M. Hoxby and K. Stange (Eds.), *Productivity in Higher Education*. National Bureau of Economic

Research.
Hoxby, C. M. and C. Avery (2013). Missing one-offs: The hidden supply of high-achieving, low-income students. Brookings Papers on Economic Activity.
Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.
Kain, J. F., D. M. O'Brien, and P. A. Jargowsky (2005). *Hopwood and the Top 10 Percent Law: How they have Affected the College Enrollment Decisions of Texas High School Graduates*. Texas School Project, University of Texas at Dallas.
Kapor, A. (2015). Distributional effects of race-blind affirmative action. Working Paper.
Kirkebøen, L., E. Leuven, and M. Mogstad (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics 131*(3), 1057–1111.
Kobrin, J. L., B. F. Patterson, E. J. Shaw, K. D. Mattern, and S. M. Barbuti (2008). Validity of the SAT® for predicting first-year college grade point average. research report no. 2008-5. Technical report, College Board.
Long, M. C. (2004). Race and college admissions: An alternative to affirmative action? *Review of Economics and Statistics 86*(4), 1020–1033.
MacLeod, W. B., E. Riehl, J. E. Saavedra, and M. Urquiola (2017, July). The big sort: College reputation and labor market outcomes. *American Economic Journal: Applied Economics 9*(3), 223–261.
MacLeod, W. B. and M. Urquiola (2015). Reputation and school competition. *American Economic Review 105*(11), 3471–3488.
Mogstad, M., A. Torgovitsky, and C. R. Walters (2019). Identification of causal effects with multiple instruments: Problems and some solutions. National Bureau of Economic Research Working Paper No. 25691.
Niu, S. X. and M. Tienda (2010). The impact of the Texas Top Ten Percent Law on college enrollment: A regression discontinuity approach. *Journal of Policy Analysis and Management 29*(1), 84–110.
Ramey, G. and V. A. Ramey (2010). The rug rat race. *Brookings Papers on Economic Activity 41*(1 (Spring)), 129–199.
Riehl, E., J. E. Saavedra, and M. Urquiola (2016). Learning and earning: An approximation to college value added in two dimensions. Chapter in *Productivity in Higher Education* (forthcoming), Caroline Hoxby and Kevin Stange, editors.
Rothstein, J. M. (2004). College performance predictions and the SAT. *Journal of Econometrics 121*(1), 297–317.
Saavedra, J. E. (2009). The returns to college quality: A regression discontinuity analysis. Harvard University.
Smith, J., M. Pender, and J. Howell (2013). The full extent of student-college academic undermatch. *Economics of Education Review 32*, 247–261.
Wager, S. and S. Athey (2017). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association (forthcoming).
World Bank (2003). Tertiary education in colombia: Paving the way for reform. Technical report, The World Bank.
Zimmerman, S. (2018). Elite colleges and upward mobility to top jobs and top incomes. *American Economic Review* (forthcoming).

Figures

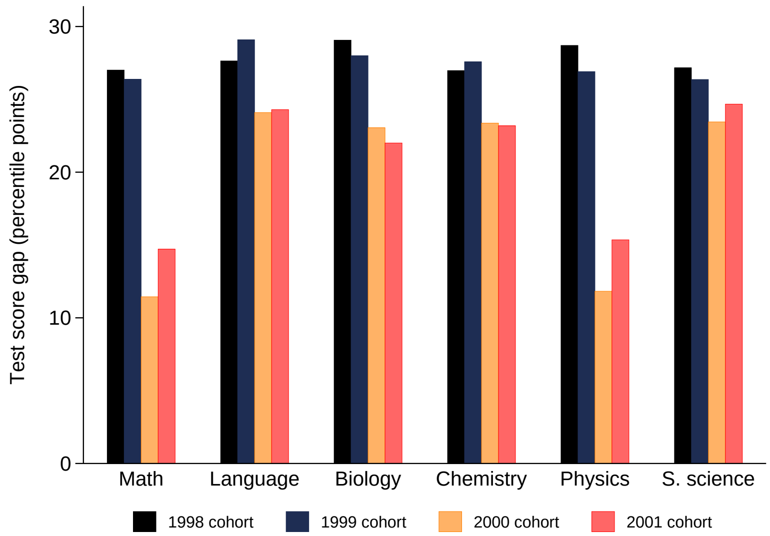
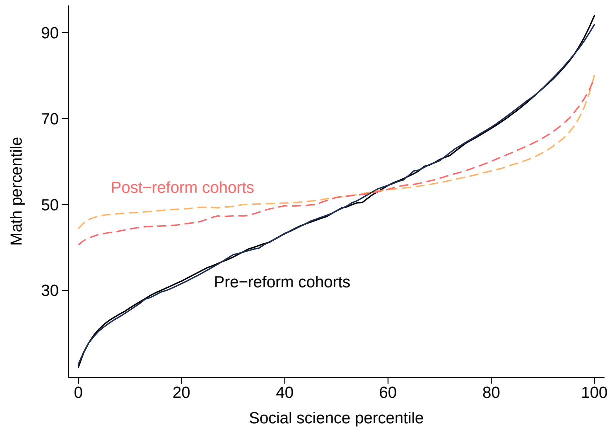
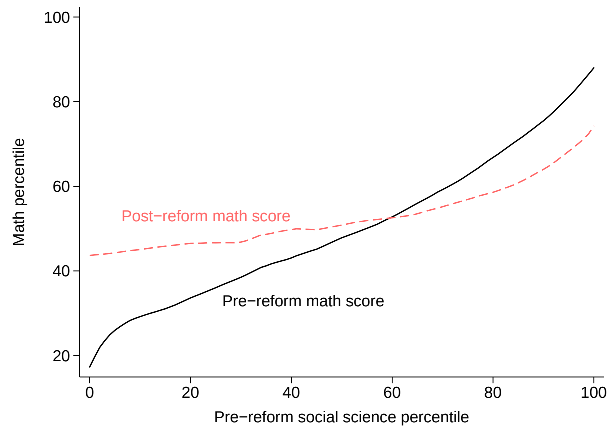


Figure 1. Test score gaps by family income (*Top quartile* − *Bottom quartile*)

Notes: The height of each bar is the test score gap by family income, defined as the difference in mean test percentile between students in the top and bottom income quartiles.



Panel A. All admission exam takers

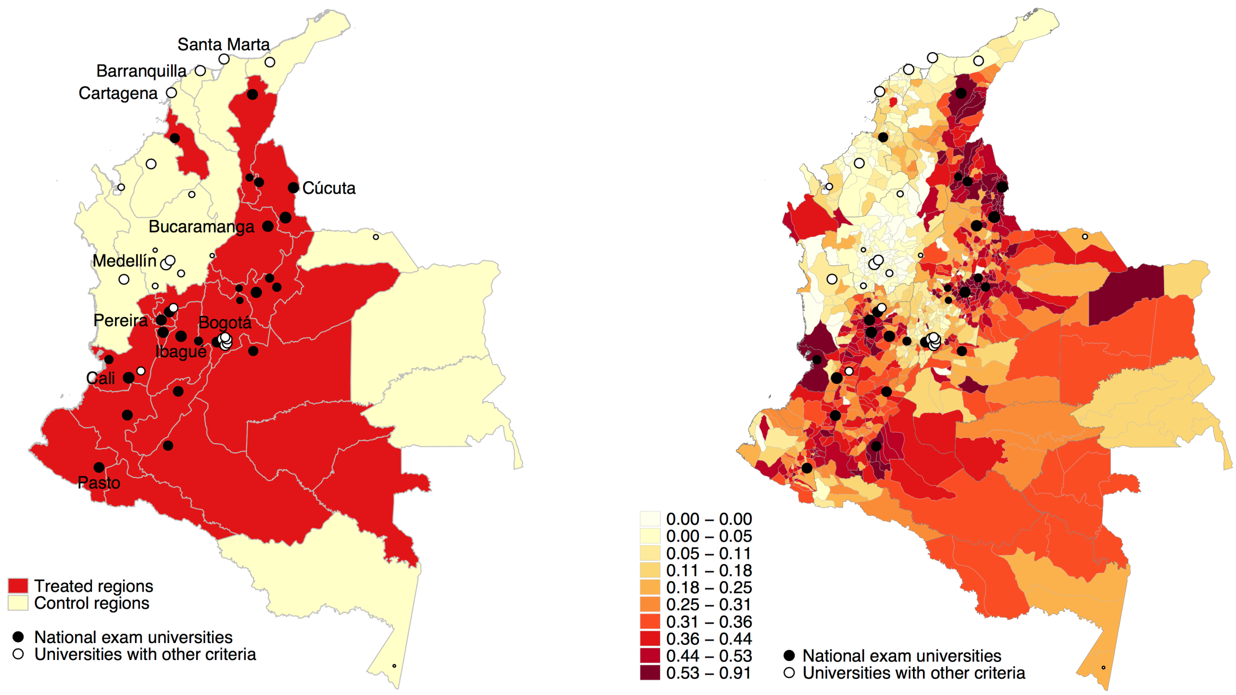


Panel B. Students who took *both* the pre- and post-reform exams

FIGURE 2. Relationship between math and social science admission scores

Notes: Panel A plots local linear regressions of admission exam math percentiles on social science percentiles for each cohort. The sample includes all students in column (A) of Table 1.

Panel B plots analogous regressions for the subset of students who took *both* the pre- and post-reform admission exams (N = 40,063). I plot estimates separately for pre- and post-reform math percentiles, but use *pre-reform* social science percentiles in both regressions. Percentiles are computed relative to students in this subsample.



Panel A. Region-level treatment

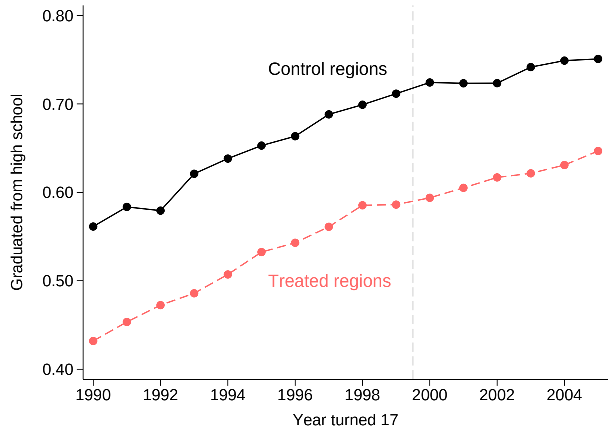
Panel B. Municipality-level treatment

FIGURE 3. Treatment variables

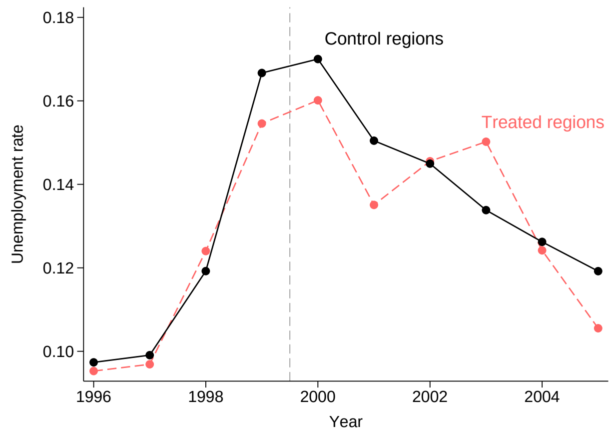
Notes: Black dots depict public universities that use only national exam scores in admissions. White dots are public universities with other admission criteria. Dot sizes are proportional to enrollment.

Panel A shows the region-level treatment variable, which is the (enrollment-weighted) modal public university admission method. I define treatment for regions with no universities using the closest public university to the capital city. Bogotá is its own administrative region, and the map does not show the island region San Andrés y Providencia. Both are control regions. See Appendix Table A2 for details.

Panel B defines treatment as the fraction of a municipality's college students from the 1998–1999 exam cohorts who enrolled in a public university with national exam admissions. Colors depict deciles of this treatment variable, with darker shades reflecting higher treatment intensity.



Panel A. High school graduation rate

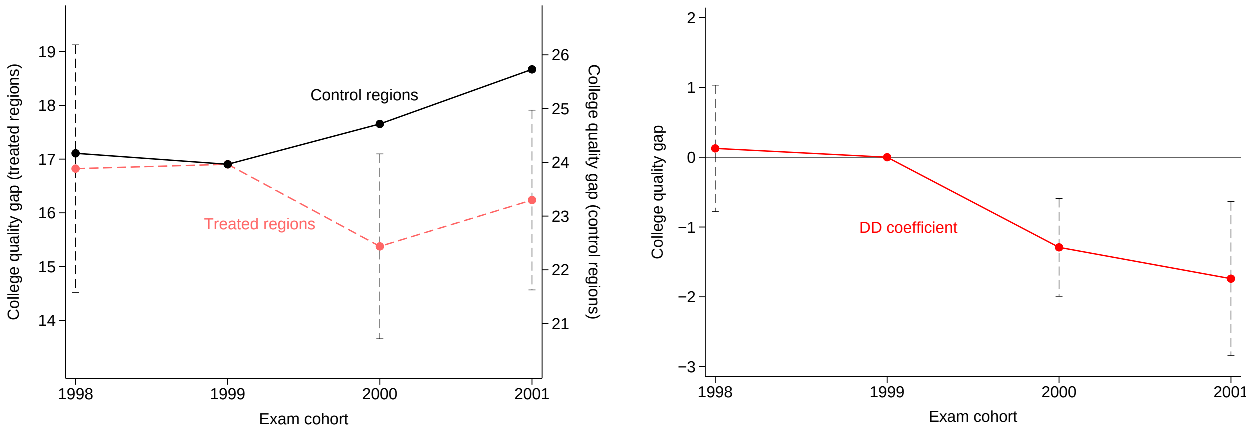


Panel B. Unemployment rate

FIGURE 4. High school graduation and unemployment rates in treated and control regions

Notes: Panel A shows the survey-weighted mean high school graduation rate in treated and control regions for cohorts defined by the year individuals turned 17. The vertical line separates cohorts who turned 17 before and after the admission exam reform. I use the 2011–2014 waves of the household survey *Gran Encuesta Integrada de Hogares*.

Panel B shows the mean unemployment rate in treated and control regions, with a vertical line separating pre- and post-reform years. Data are from the household survey *Encuesta Nacional de Hogares*. For 1996–2000, data are only available at the region level for the 24 largest regions, and the graph displays averages using population weights. For 2001–2005, I use individual level data for the same 24 regions and compute averages using survey weights.



Panel A. Treated and control regions

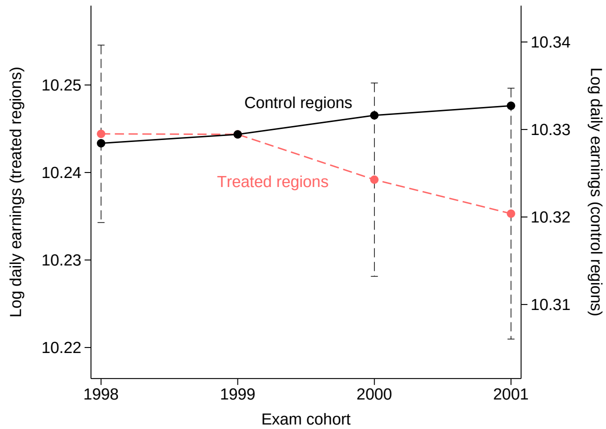
Panel B. Municipality-level event study

FIGURE 5. Reform effects on the college quality gap by family income (*Top quartile* − *Bottom quartile*)

Notes: In Panel A, the dashed line is the difference in mean college quality between top and bottom income quartile students in treated regions. The solid line is the mean college quality gap in control regions.

Panel B plots event-study estimates of the reform's effect on the family income college quality gap using the municipality-level treatment variable and municipality fixed effects. The event study interacts the θ^q coefficient from equation (8) with dummies for exam cohorts t, omitting the 1999 cohort interaction.

Dashed lines are 95% confidence intervals using standard errors from region- (Panel A) and municipality-level (Panel B) event-study regressions. Standard errors are clustered at the region level.



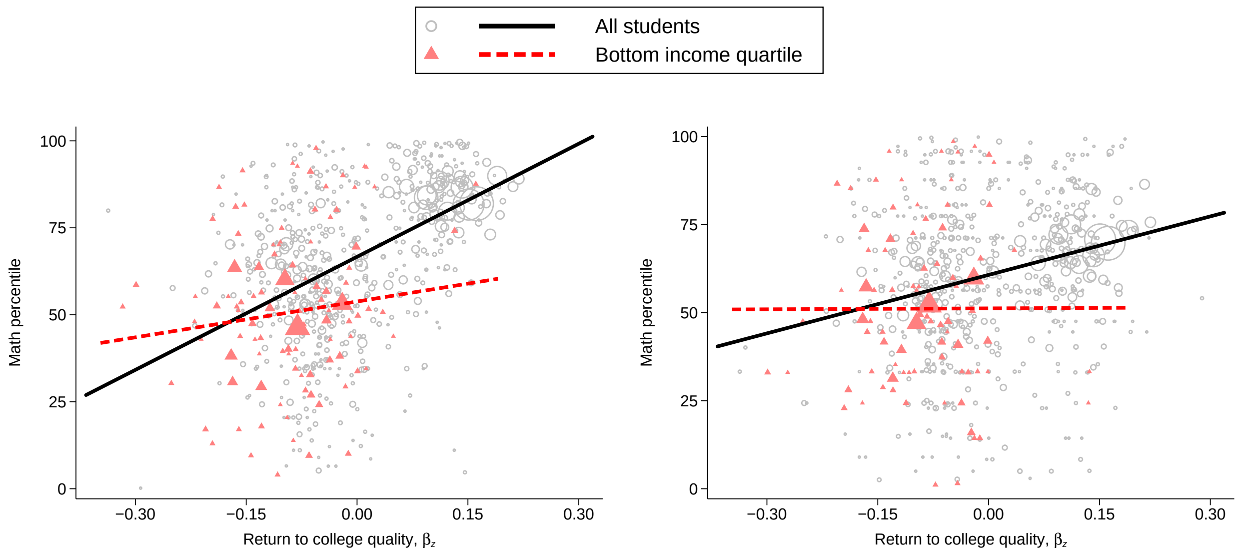
Panel A. Treated and control regions

Panel B. Municipality-level event study

FIGURE 6. Reform effects on log daily earnings (*All students*)

Notes: Panel A plots mean log daily earnings in treated regions (solid line) and control regions (dashed line). Panel B plots event-study estimates of the reform's effect on log daily earnings using municipality-level treatment. The event study interacts the θ coefficient from equation (7) with dummies for cohorts t, omitting the 1999 cohort interaction.

Dashed lines are 95% confidence intervals using standard errors from region- (Panel A) and municipality-level (Panel B) event-study regressions. Standard errors are clustered at the region level.



Panel A. Pre-reform exam cohorts

Panel B. Post-reform exam cohorts

FIGURE 7. Math admission exam scores and returns to college quality, ${}_z$

Notes: The x-axis in each panel depicts ${}_z$ estimates from Table 7, Panel B. The y-axis is the mean math exam percentile in each covariate cell z. Panel A uses pre-reform cohorts (1998–1999). Panel B uses post-reform cohorts (2000–2001). The sample includes cells z with first stage F statistics above ten (Table 7, Panel B, column (D)).

Triangles represent students in the bottom income quartile. Hollow circles represent all other students. Marker sizes are proportional to covariate cell size. The solid black line is the linear relationship between math scores and ${}_z$ for all students. The red dashed line is the linear relationship between these variables for bottom income quartile students. Both lines are from covariate cell level regressions with observations weighted by cell size.

Tables

Table 1. Summary statistics

	(A)	(B)	(C)	(D)	(E)	(F)
	Total # of students		Means for pre-reform cohorts (1998–1999)			
Characteristic	High school graduates	College enrollees	Admit exam math score (percentile)	College quality (percentile)	Graduated from college	Daily earnings (2009 USD)
Top income quartile	223,426	152,358	70.4	62.9	0.48	17.45
Bottom income quartile	657,295	154,261	44.1	42.5	0.34	12.14
College educated mother	228,009	152,634	68.6	61.8	0.48	17.03
Primary educated mother	936,043	249,904	46.1	44.7	0.36	12.64
High ranked high school	315,699	200,771	69.3	63.1	0.48	16.82
Low ranked high school	841,124	203,277	42.9	38.9	0.31	12.00
All students	1,644,260	612,949	51.2	50.6	0.39	14.13

Notes: Column (A) includes students who took the national college admission exam in 1998–2001. Column (B) includes the subset of these students who enrolled in college. Columns (C)–(F) display means for the 1998–1999 exam cohorts, with column (C) including all exam takers and columns (D)–(F) including only college enrollees. See Appendix C.1 for details on the sample and variable definitions.

Table 2. Reform effects on test score gaps

	(A)	(B)	(C)	(D)	(E)	(F)	(G)
		Dependent variable: Admission exam score (percentile)					
		Reform effect by subject					
	Pre-reform mean gap	Math	Lang	Biol	Chem	Phys	S. sci
Family income gap	27.6	-13.7^{***}	-4.2^{***}	-6.0^{***}	-4.0^{***}	-14.3^{***}	-2.7^{***}
Top Q − Bottom Q		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Mother's education gap	24.2	-12.0^{***}	-2.8^{***}	-4.1^{***}	-2.8^{***}	-12.2^{***}	-2.1^{***}
College − Primary		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
High school rank gap	30.7	-17.1^{***}	-3.2^{***}	-5.2^{***}	-4.0^{***}	-17.7^{***}	-1.3^{***}
High − Low		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)

Notes: Column (A) shows mean test score gaps in percentile points across the six exam subjects for the 1998–1999 cohorts. (These gaps were similar in each exam subject.) For columns (B)-(G), I regress test scores in each subject on cohort dummies, an indicator for the high SES group, and the interaction of this indicator with a dummy for the 2000–2001 cohorts. Columns (B)-(G) display the coefficient on the interaction variable.

The sample for this table includes students in Column (A) of Table 1. Parentheses contain robust standard errors.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 3. Reform effects on test validity

	(A)	(B)	(C)	(D)
	Dependent variable (in st. deviation units)			
	Exit exam score	Graduated from college	First year GPA	S. science exam score
Math	-0.176^{***}	-0.058^{***}	-0.129^{**}	-0.415^{***}
	(0.004)	(0.003)	(0.061)	(0.001)
Language	-0.088^{***}	0.009^{***}	-0.003	-0.211^{***}
	(0.004)	(0.003)	(0.052)	(0.001)
Biology	-0.039^{***}	0.009^{***}	-0.050	-0.258^{***}
	(0.004)	(0.003)	(0.057)	(0.001)
Chemistry	0.013^{***}	-0.003	-0.091	-0.169^{***}
	(0.004)	(0.003)	(0.059)	(0.001)
Physics	-0.148^{***}	-0.063^{***}	-0.109^{*}	-0.453^{***}
	(0.004)	(0.003)	(0.059)	(0.001)
Social science	-0.008^{**}	0.034^{***}	-0.018	
	(0.004)	(0.003)	(0.063)	
N	242,887	612,949	3,083	1,644,260
Mean pre-reform corr.	0.386	0.112	0.093	0.731

Notes: Each coefficient is from a separate regression of the dependent variable in the column header on cohort dummies, the admission score, and the interaction of the admission score with a dummy for the 2000–2001 cohorts. The table displays coefficients on the interaction term. The sample for column (C) includes students in the 2000–2004 entering cohorts at a public flagship university for whom I have transcript data (see Appendix C.1). All covariates in column (C) regressions are interacted with dummies for the number of years since the admission exam.

In all regressions, variables are normalized to standard deviation one within each exam cohort so that coefficients can be interpreted as correlation coefficients. In columns (A)–(C), admission exam scores and dependent variables are residuals from regressing these variables on cells defined by a student's college and exam cohort.

The bottom row shows the mean correlation between the dependent variable and the six exam subject scores in the 1998–1999 cohorts. Parentheses contain robust standard errors.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4. Treated and control region characteristics and balance tests

	(A)	(B)	(C)	(D)	(E)	(F)
	Pre-reform means		Region-level balance tests		Municipality-level balance tests	
Dependent variable	Treated regions	Control regions	Coef	SE	Coef	SE
Panel A. Student characteristics						
Top income quartile	0.227	0.279	−0.002	(0.003)	0.000	(0.001)
Bottom income quartile	0.264	0.235	0.001	(0.004)	−0.001	(0.002)
Mother's years of education	7.161	7.529	-0.087^{*}	(0.045)	-0.035^{**}	(0.016)
Father's years of education	7.455	7.970	−0.081	(0.062)	−0.033	(0.023)
Private high school	0.321	0.429	0.000	(0.009)	0.003	(0.003)
High ranked high school	0.181	0.209	−0.005	(0.005)	−0.001	(0.002)
Combined index (log earnings)	10.512	10.535	−0.000	(0.002)	0.001	(0.001)
Panel B. Labor market conditions						
Labor force participation rate	0.604	0.579	−0.001	(0.009)		
Unemployment rate	0.141	0.146	−0.006	(0.011)		
Panel C. Exam taking and college enrollment						
Annual exam takers (1000s)	11.545	16.474	0.508	(1.261)	0.014	(0.012)
Retook exam within one year	0.102	0.055	−0.010	(0.013)	−0.006	(0.005)
Enrolled in any college	0.343	0.362	−0.003	(0.009)	−0.000	(0.004)
Years between HS and college	3.567	3.372	−0.145	(0.087)	−0.059	(0.047)
Kilometers from HS to college	102.925	59.077	−3.648	(3.256)	−2.393	(1.561)
Stayed in region if enrolled	0.572	0.850	0.016^{*}	(0.008)	0.009^{*}	(0.005)
Public university if enrolled	0.468	0.275	0.030^{*}	(0.016)	0.009	(0.006)
Regions	20	13				

Notes: The sample for Panel A includes all students in column (A) of Table 1. Income quartiles in Panel A are defined relative to all exam takers in each cohort. I compute the combined index by regressing 2012 log daily earnings on all covariates in Appendix Table C5 using only pre-reform cohorts. The index is the predicted value from this regression in all cohorts. The data for Panel B are from the Colombian labor market survey *Encuesta Nacional de Hogares*. These data contain region-level statistics from the 24 largest regions. The sample for Panel C is the same as in Panel A, except the last four rows include only students who enrolled in college.

Columns (A)–(B) present means of each variable for 1998–1999 exam takers in treated and control regions. Columns (C)–(D) display estimates of θ and its standard error from equation (7) using the region-level treatment variable and region fixed effects. Columns (E)–(F) display analogous estimates using the municipality-level treatment variable and municipality fixed effects. Standard errors are clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5. Reform effects on SES college quality gaps

	(A)	(B)	(C)	(D)	(E)	(F)
	Dependent variable: College quality, Q_c					
	Benchmark specification (8)		Within population groups		Within population $\times$ proximity groups	
	Region-level	Muni-level	Region-level	Muni-level	Region-level	Muni-level
Family income gap	-2.40^{***}	-1.56^{***}	-3.35^{***}	-1.61^{***}	-3.52^{***}	-1.89^{***}
Top Q − Bottom Q	(0.54)	(0.36)	(0.93)	(0.46)	(1.05)	(0.56)
Mother's education gap	-1.41^{***}	-0.80^{**}	-1.18^{*}	-0.58	-1.51^{*}	-0.76
College − Primary	(0.49)	(0.31)	(0.67)	(0.40)	(0.76)	(0.47)
High school rank gap	-1.48	-1.31^{**}	-1.09	-0.99^{*}	-1.01	-0.97
High − Low	(0.92)	(0.48)	(1.07)	(0.52)	(1.22)	(0.66)

Notes: This table displays θ^q coefficients from separate regressions (8). The sample includes students in Column (B) of Table 1. Rows are defined by the three SES measures from Table 1.

Columns (A) reports estimates from equation (8) using the region-level treatment variable (Treatment_g) and fixed effects for regions ($_g$ and $_{gt}$). Column (B) uses the municipality-level treatment variable and fixed effects for municipalities. Columns (C)–(D) are analogous to columns (A)–(B), but regressions include dummies for eleven municipality population groups fully interacted with dummies for each exam cohort and SES group. Columns (E)–(F) include dummies for full interactions between municipality population groups, municipality proximity groups, exam cohorts, and SES groups. See the text and Appendix Table A6 for details on the municipality groups.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6. Reform effects on college persistence, graduation, and earnings

	(A)	(B)	(C)	(D)	(E)	(F)
	Dep. variable: One year persistence		Dep. variable: College graduation		Dep. variable: Log daily earnings	
	Region-level	Muni-level	Region-level	Muni-level	Region-level	Muni-level
All students	-0.014^{***}	-0.010^{***}	-0.015^{*}	-0.008^{**}	-0.013^{**}	-0.007^{**}
	(0.005)	(0.003)	(0.008)	(0.004)	(0.005)	(0.003)
Top income quartile	-0.017^{*}	-0.012^{***}	-0.027^{***}	-0.011^{**}	-0.018^{*}	-0.008
	(0.009)	(0.004)	(0.009)	(0.005)	(0.009)	(0.006)
Bottom income quartile	-0.021^{***}	-0.013^{***}	-0.010	-0.008	-0.013^{***}	-0.008^{**}
	(0.006)	(0.003)	(0.009)	(0.005)	(0.005)	(0.003)
College educated mother	-0.016^{*}	-0.010^{**}	-0.022^{**}	-0.008	-0.011	-0.004
	(0.008)	(0.004)	(0.010)	(0.006)	(0.011)	(0.006)
Primary educated mother	-0.013^{**}	-0.010^{***}	-0.007	-0.007	-0.010^{**}	-0.007^{***}
	(0.005)	(0.003)	(0.008)	(0.004)	(0.005)	(0.002)
High ranked high school	-0.015^{**}	-0.010^{**}	-0.019^{**}	-0.011^{**}	-0.006	-0.006
	(0.007)	(0.004)	(0.007)	(0.005)	(0.009)	(0.005)
Low ranked high school	-0.014^{*}	-0.008^{**}	-0.010	-0.004	-0.013^{**}	-0.005
	(0.007)	(0.003)	(0.011)	(0.005)	(0.005)	(0.003)

Notes: This table displays θ coefficients from separate regressions (7) with three dependent variables: an indicator equal to one if the student was still in college one year after enrolling (columns (A)–(B)); an indicator for college graduation (columns (C)–(D)); and log daily earnings 10–11 years after the admission exam (columns (E)–(F)). The sample in columns (A)–(D) includes students in Column (B) of Table 1; columns (E)–(F) include only students with formal sector earnings. The first row shows estimates for all students; other rows report estimates by SES group.

Columns (A), (C), and (E) report estimates using the region-level treatment variable, Treatment_g, and region dummies, $\ _g$. Columns (B), (D), and (F) use the municipality-level treatment variable and municipality dummies.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7. IV estimates of returns to college quality, $_{z}$

Panel A. Average returns by SES (standard 2SLS estimation)

	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
		First stage effects, θ_z			Returns to college quality, $_{z}$			
	N	Coef	SE	F stat	Coef	SE	# coefs	St. dev
Top income quartile	84,615	-0.196	(0.052)	14.2	0.091	(0.055)	1	.
Bottom income quartile	83,587	0.152	(0.082)	3.4	-0.087	(0.040)	1	.
College educated mother	85,604	-0.100	(0.057)	3.1	0.111	(0.127)	1	.
Primary educated mother	138,655	0.116	(0.081)	2.1	-0.084	(0.056)	1	.
High ranked high school	118,394	-0.066	(0.073)	0.8	0.090	(0.147)	1	.
Low ranked high school	102,874	0.112	(0.074)	2.3	-0.112	(0.085)	1	.
All students	340,623	0.024	(0.074)	0.1	-0.526	(1.529)	1	.

Panel B. Across- and within-SES returns (IV random forest estimation)

		First stage effects, θ_z			Returns to college quality, $_{z}$			
	N	Mean coef	Mean SE	N with $F > 10$	Mean coef	Mean SE	# coefs	St. dev
Top income quartile	35,287	-0.135	0.118	2,196	0.125	0.074	273	0.041
Bottom income quartile	31,817	0.074	0.095	516	-0.087	0.101	127	0.066
College educated mother	36,164	-0.094	0.110	1,792	0.125	0.075	182	0.052
Primary educated mother	50,940	0.062	0.096	624	-0.069	0.099	176	0.093
High ranked high school	48,883	-0.076	0.110	2,234	0.124	0.073	289	0.041
Low ranked high school	39,159	0.095	0.092	1,000	-0.098	0.097	265	0.055
All students	135,200	0.022	0.101	4,548	0.025	0.081	947	0.109

Notes: Panel A shows estimates from equations (9)–(10) using region-level treatment and covariate groups z defined by the SES group in the first column. Column (A) shows the sample size for each regression, which includes students in column (B) of Table 1 with formal sector earnings. Columns (B)–(C) report estimates of θ_z and its standard error from equation (9). Column (D) shows F statistics from tests of $\theta_z = 0$. Columns (E)–(F) report estimates of $_{z}$ and its standard error from equation (10). Column (G) shows that there is one coefficient per SES group (in contrast to Panel B). Standard errors in columns (C) and (F) are clustered at the region level.

Panel B summarizes Athey et al. (2017) random forest estimates from equations (9)–(10) (see Appendix C.5). The estimation uses a 50 percent training sample, and this table presents statistics from the validation sample. Column (A) shows the number of students in the validation sample, which excludes students with missing covariates, Z_i. Columns (B)–(C) show the mean of the θ_z estimates and the average standard error. Column (D) shows the number of students in the validation sample with first stage F statistics above ten; columns (E)–(H) present statistics only from this subsample. Columns (E)–(F) show the mean of the $_{z}$ estimates and the average standard error. Column (G) shows the number of unique $_{z}$ parameters. Column (H) reports the standard deviation of the $_{z}$ estimates.

In all regressions, I divide Q_c by ten so that one unit equals ten percentile points in the distribution of colleges.

Table 8. Reform effects on the correlation of returns to college quality, $_z$, and admission scores

	(A)	(B)	(C)	(D)
	Dependent variable: Return to college quality, $_z$			
	All students		Bottom income quartile	
Exam subject	Pre-reform correlation	Reform effect	Pre-reform correlation	Reform effect
Math	0.427***	-0.212***	0.109	-0.103
	(0.027)	(0.037)	(0.104)	(0.169)
Language	0.413***	-0.022	0.069	-0.002
	(0.027)	(0.036)	(0.100)	(0.172)
Biology	0.411***	-0.033	0.067	0.031
	(0.026)	(0.036)	(0.104)	(0.172)
Chemistry	0.433***	-0.008	0.122	-0.111
	(0.026)	(0.036)	(0.104)	(0.170)
Physics	0.423***	-0.172***	0.057	-0.106
	(0.026)	(0.036)	(0.106)	(0.175)
Social science	0.398***	-0.007	0.028	-0.016
	(0.027)	(0.036)	(0.103)	(0.175)
N	2,034	4,548	294	516

Notes: The sample includes students from column (D) in Table 7 (Panel B). Column (A) reports coefficients from separate regressions of random forest returns, $_z$, on admission scores for each subject using the 1998–1999 cohorts. Column (B) shows the change in these coefficients between the 2000–2001 and 1998–1999 cohorts. Columns (C)–(D) are analogous, but regressions include only bottom income quartile students. All variables are normalized to standard deviation one within the pre- and post-reform cohorts so coefficients can be interpreted as bivariate correlations.

Estimates of $_z$ use the region-level treatment variable; Appendix Table A18 replicates this table with municipality-level estimates of $_z$. Parentheses contain robust standard errors adjusted for estimation error. For this I form a diagonal matrix where each element is the squared residual from regressing $_z$ on the admission score plus the squared standard error of $_z$. I then use this diagonal error matrix to compute robust standard errors.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix — For Online Publication

A. Appendix figures and tables

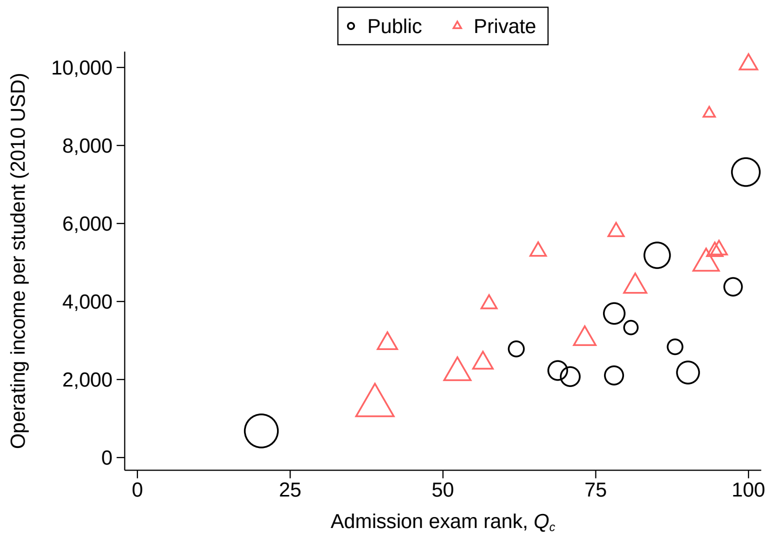
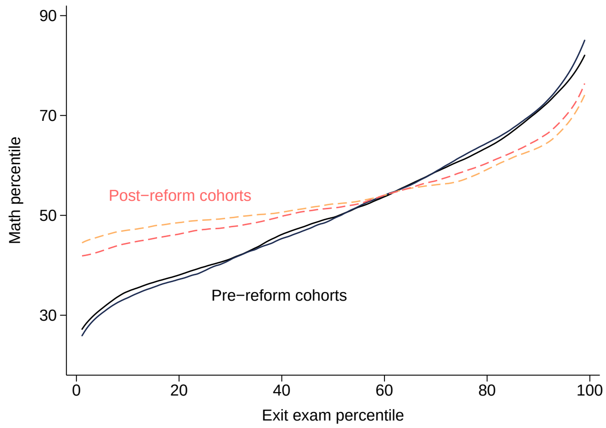


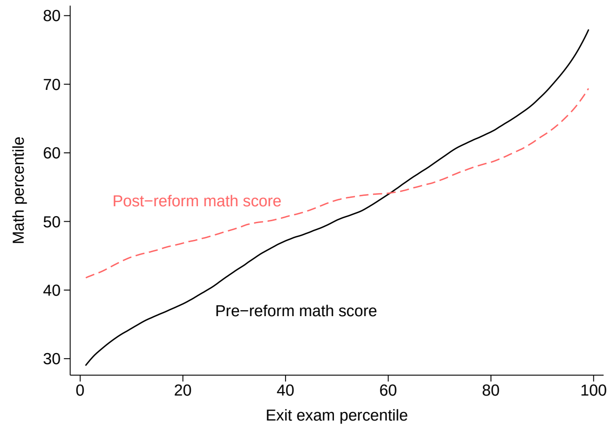
Figure A1. Admission exam rank, Q_c and operating income per student for 26 colleges

Notes: The x-axis is my main college quality measure Q_c. The y-axis depicts operating income per student in 2010 U.S. dollars. Income data are from a Ministry of Education list of the top 26 colleges by 2010 operating income (available in May 2018 at: https://www.mineducacion.gov.co/observatorio/1722/article-280538.html). I compute the number of students at each university for the year 2013 from the Ministry of Education's administrative enrollment records used elsewhere in this paper. The enrollment-weighted correlation between these two variables is 0.77. The unweighted correlated is 0.66.

Black circles are public colleges; red triangles are private colleges. Marker sizes are proportional to the number of students.



Panel A. All exit exam takers



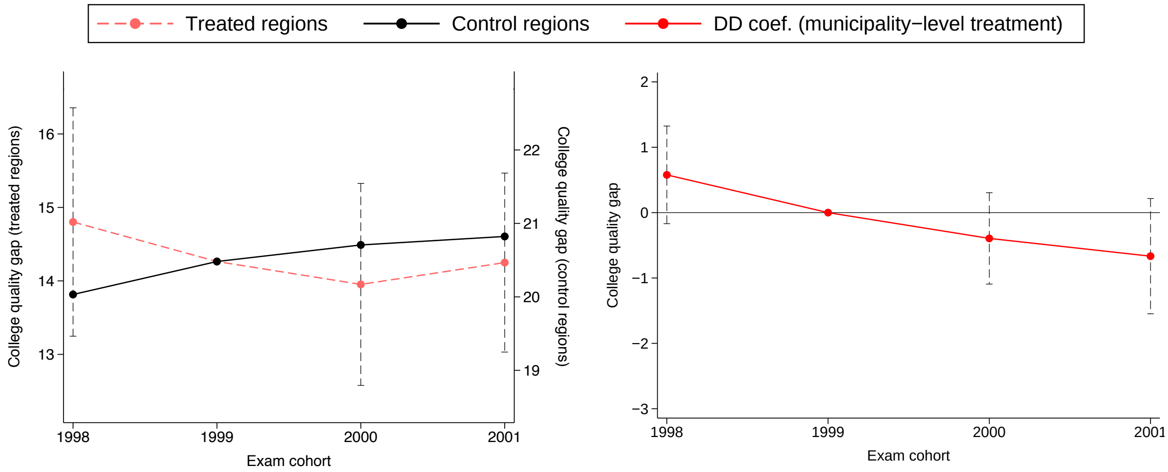
Panel B. Exit exam takers who took *both* admission exams

FIGURE A2. Relationship between math admission scores and exit exam scores

Notes: This figure is analogous to Figure 2, but the it uses scores on a field-specific college exit exam—rather than social science admission scores—as a measure of individual ability. Exit exam scores are not a perfect measure because they are confounded by the admission exam's effects on students' college choices. Nonetheless, the dramatic reductions in validity suggest that selection issues are likely to be a second order effect relative to changes in the abilities measured by the exams.

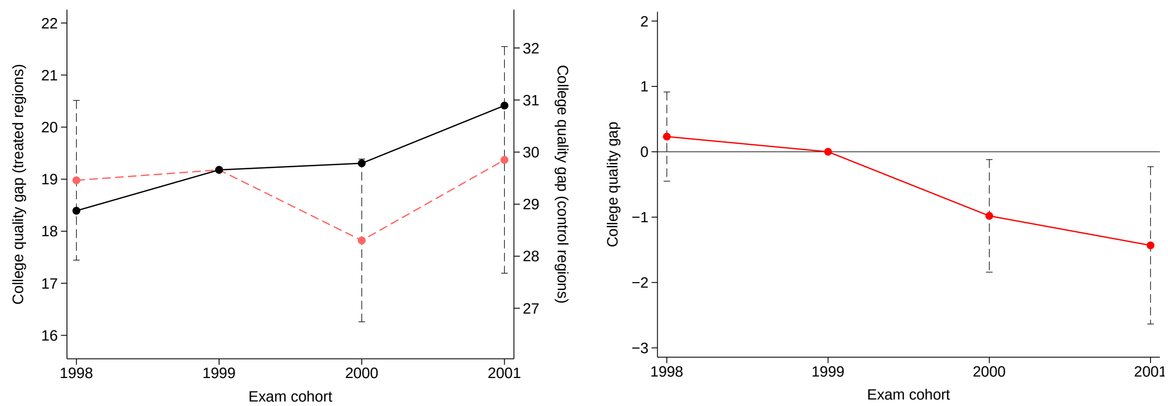
Panel A plots local linear regressions of admission exam math percentiles on exit exam percentiles for each cohort. The sample includes all students who took the exit exam ($N = 281{,}473$). Percentiles are computed relative to students in this subsample.

Panel B plots analogous regressions for the subset of exit exam takers who took *both* the pre- and post-reform admission exams ($N = 10{,}370$). I plot estimates separately for pre- and post-reform math percentiles. Percentiles are computed relative to students in this subsample.



Panel A. Mother's education gap
(*Treated and control regions*)

Panel B. Mother's education gap
(*Municipality-level event study*)

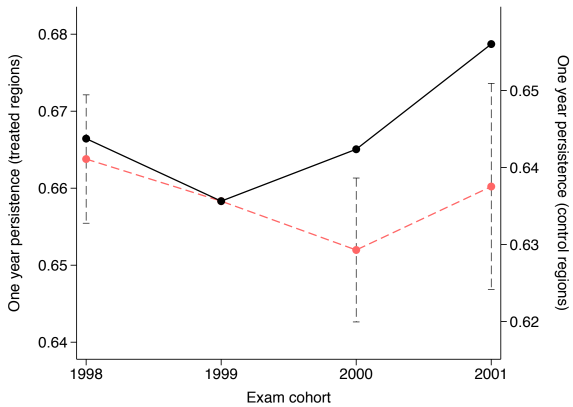


Panel C. High school rank gap
(*Treated and control regions*)

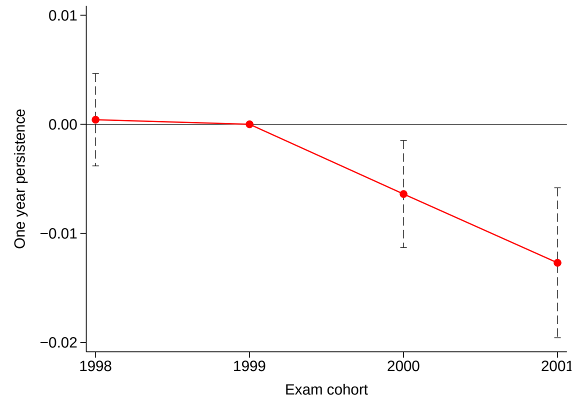
Panel D. High school rank gap
(*Municipality-level event study*)

FIGURE A3. Reform effects on SES college quality gaps

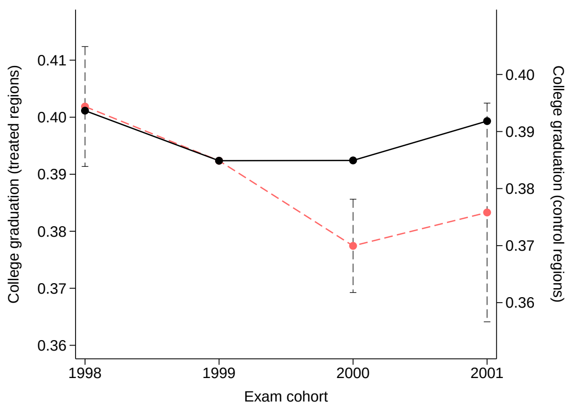
Notes: This figure plots region- and municipality-level estimates of the reform's effect on SES college quality gaps using the mother's education and high school rank measures of socioeconomic inequality defined in Table 1. Panels A and C are analogous to Panel A in Figure 5. Panels B and D are analogous to Panel B in Figure 5. See the notes to Figure 5 for details.



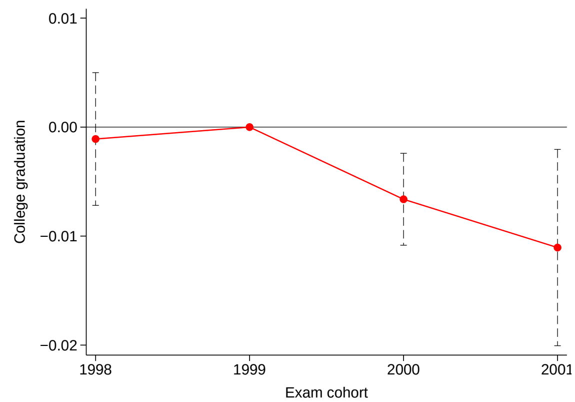
Panel A. One year persistence
(*Treated and control regions*)



Panel B. One year persistence
(*Municipality-level event study*)



Panel C. College graduation
(*Treated and control regions*)



Panel D. College graduation
(*Municipality-level event study*)

FIGURE A4. Reform effects on graduation and persistence

Notes: This figure plots region- and municipality-level estimates of the reform's effect on one year college persistence (Panels A–B) and college graduation (Panels C–D). Panels A and C are analogous to Panel A in Figure 6. Panels B and D are analogous to Panel B in Figure 6. See the notes to Figure 6 for details.

Table A1. Admission exam rank, Q_c, and other college characteristics by institution tier

(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
Institutional level	College ownership	Total enrollment	Mean years to grad	Admission exam rank (Q_c)	Admit rate	Grad rate	Annual tuition (2012 USD)
University	Private	484,487	5.6	55.0	0.82	0.45	2,877
University	Public	543,064	5.6	58.7	0.47	0.42	440
University Institute	Private	300,852	4.4	29.5	0.86	0.39	1,270
University Institute	Public	95,063	4.0	19.0	0.72	0.36	472
Technology Institute	Private	51,307	3.1	29.1	0.90	0.26	950
Technology Institute	Public	35,242	3.5	46.9	0.85	0.39	799
Technical/Professional Institute	Private	40,173	3.1	22.7	0.83	0.24	830
Technical/Professional Institute	Public	6,787	2.8	5.5	0.93	0.39	415

Notes: This table shows how my college quality measure, Q_c, relates to other institutional characteristics.

Columns (A) and (B) categorize colleges into tiers based on their institution level and ownership status. Technology Institutes also include the Ministry of Education's fifth category of college, Technology Schools.

Column (C) shows total enrollment in each tier for the year 2013. Column (D) shows average time to graduation. Column (E) shows the average value of my main college quality measure, Q_c. Column (G) shows the average graduation rate in the 2002–2003 enrollment cohorts. Each of these statistics is computed from the Ministry of Education's administrative enrollment records used elsewhere in this paper.

Column (F) shows the average admission rate (number admitted/number applied) over the years 2007–2013 using college-program-cohort level data from the Ministry of Education (available in May 2018 at: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-212400.html). Column (H) shows the median annual tuition rate in each tier for 2012 as reported by the Ministry of Education (available in May 2018 at: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-212350_resumen.xls). Within each institution tier, I compute a weighted average of the median tuitions for each of three "program levels" (technical professional, technical, and university) using the number of students who enrolled in each program level as weights.

Table A2. Public universities and definition of treated regions

(A) Region	(B) Public university	(C) Cohort size	(D) Admission method	(E) Closest region	(F) Treated region?
Amazonas	Univ. Nacional de Colombia (Leticia)	28	Other		
Antioquia	Univ. de Antioquia (Medellin)	6,962	Other		
	Univ. Nacional de Colombia (Medellin)	2,253	Other		
	Univ. de Antioquia (Carmen De Viboral)	220	Other		
	Univ. de Antioquia (Turbo)	165	Other		
	Univ. de Antioquia (Caucasia)	128	Other		
	Univ. de Antioquia (Andes)	101	Other		
	Univ. de Antioquia (Puerto Berrio)	50	Other		
	Univ. de Antioquia (Antioquia)	49	Other		
Arauca	Univ. Nacional de Colombia (Arauca)	65	Other		
Atlantico	Univ. del Atlantico	3,609	Other		
Bogota	Univ. Nacional Abierta y A Distancia	14,589	Other		
	Univ. Distrital Francisco Jose de Caldas	5,095	National exam		
	Univ. Nacional de Colombia (Bogota)	4,882	Other		
	Univ. Ped. Nacional	1,910	Other		
	Univ. Militarnueva Granada	1,896	Other		
	Univ. Colegio Mayor de Cundinamarca	1,365	Other		
Bolivar	Univ. de Cartagena	2,978	Other		
Boyaca	Univ. Ped. y Tecn. de Colombia (Tunja)	3,196	National exam		✓
	Univ. Ped. y Tecn. de Colombia (Sogamoso)	548	National exam		
	Univ. Ped. y Tecn. de Colombia (Duitama)	490	National exam		
	Univ. Ped. y Tecn. de Colombia (Chiquinquira)	150	National exam		
Caldas	Univ. de Caldas	2,810	National exam		✓
	Univ. Nacional de Colombia (Manizales)	991	Other		
Caqueta	Univ. de la Amazonia	1,274	National exam		✓
Casanare		431		Boyaca	✓
Cauca	Univ. del Cauca	2,221	National exam		✓
Cesar	Univ. Popular del Cesar (Valledupar)	2,488	National exam		✓
	Univ. Popular del Cesar (Aguachica)	233	National exam		
Choco	Univ. Tecn. del Chocodiego Luis Cordoba	2,278	Other		
Cordoba	Univ. de Cordoba	2,210	Other		
Cundinamarca	Univ. de Cundinamarca (Fusagasuga)	2,013	National exam		✓
	Univ. de Cundinamarca (Girardot)	421	National exam		
	Univ. de Cundinamarca (Ubate)	134	National exam		
Guainia		8		Arauca	
Guaviare		58		Meta	✓
Huila	Univ. Surcolombiana	1,750	National exam		✓

Notes: This table is continued on the next page.

Table A2. Public universities and definition of treated regions (continued)

(A) Region	(B) Public university	(C) Cohort size	(D) Admission method	(E) Closest region	(F) Treated region?
La Guajira	Univ. de la Guajira	1,970	Other		
Magdalena	Univ. del Magdalena	3,257	Other		
Meta	Univ. de Los Llanos	1,148	National exam		✓
Narino	Univ. de Narino	1,996	National exam		✓
Norte Santander	Univ. de Pamplona	6,192	National exam		✓
	Univ. Francisco de Paula Santander (Cucuta)	3,387	National exam		
	Univ. Francisco de Paula Santander (Ocana)	944	National exam		
Putumayo		260		Narino	✓
Quindio	Univ. del Quindio	3,118	National exam		✓
Risaralda	Univ. Tecn. de Pereira	2,701	National exam		✓
San Andres		102		Bolivar	
Santander	Univ. Industrial de Santander	3,838	National exam		✓
Sucre	Univ. de Sucre	1,033	National exam		✓
Tolima	Univ. del Tolima	4,099	National exam		✓
Valle del Cauca	Univ. del Valle	5,660	National exam		✓
	Univ. del Pacifico	539	National exam		
	Univ. Nacional de Colombia (Palmira)	517	Other		
Vaupes		12		Meta	✓
Vichada		27		Arauca	

Notes: This table describes my definition of treated and control regions, which is based on pre-reform admission methods in public universities.

Column (A) lists the 33 Colombian administrative regions (*departamentos*).

Column (B) lists public universities in each region; blank cells indicate that the region has no public universities.

Column (C) shows the mean number of entering students per cohort over all cohorts in my records. For regions without public universities, it shows the average number of public university enrollees per pre-reform exam cohort.

Column (D) lists the pre-reform admission method of each university. I collected information on admission methods prior to 2000 by searching through historical student regulations at each college, or by tracking down information from historical college or newspaper websites using the website archive.org. "National exam" means that the university used only national exam scores for admission. "Other" means that the university used other admission criteria; most commonly this meant that applicants were required to take the university's own admission exam, but in some cases schools considered other information such as high school GPA or personal interviews.

Column (E) shows the closest region with a public university for those regions with blank cells in column (B). Closest is defined by distance between capital cities.

Column (F) shows my classification of regions as treated or control. Treated regions are those where public universities used only the national exam for admissions. Control regions are those with public universities that use other admission methods. In the three regions with mixed admission criteria (Bogotá, Caldas, and Valle del Cauca), I define treatment as the modal admission method (weighted by column (C)). For regions with no public universities, I define treatment using the region listed in column (E).

Table A3. Reform effects on SES major selectivity gaps

	(A)	(B)	(C)	(D)	(E)	(F)
	Dependent variable: Major selectivity					
	Benchmark specification (8)		Within population groups		Within population × proximity groups	
	Region-level	Muni-level	Region-level	Muni-level	Region-level	Muni-level
Family income gap	0.37	−0.36	0.81	−0.13	0.88	−0.00
Top Q − Bottom Q	(0.81)	(0.40)	(0.89)	(0.43)	(0.92)	(0.55)
Mother's education gap	0.62	−0.00	0.96	0.15	0.82	0.13
College − Primary	(0.65)	(0.32)	(0.74)	(0.39)	(0.82)	(0.48)
High school rank gap	1.21**	0.15	0.97	0.09	0.94	−0.02
High − Low	(0.59)	(0.40)	(0.96)	(0.49)	(1.00)	(0.55)

Notes: This table is identical to Table 5 except that regressions use a different dependent variable. The dependent variable in Table 5 is college quality, Q_c, defined as a college's percentile rank based on mean pre-reform national exam scores. In this table, the dependent variable is "major selectivity," defined as a major's percentile rank based on mean pre-reform national exam scores. Majors are defined as the Ministry of Education's 53 program groups, which combine programs offered by different colleges into common categories such as economics or mechanical engineering. The estimates thus show the reform's effect on the SES gap in major selectivity.

This table displays θ^q coefficients from separate regressions (8) using the major selectivity dependent variable. The sample includes students in Column (B) of Table 1. Rows are defined by the three SES measures from Table 1.

Columns (A) reports estimates from equation (8) using the region-level treatment variable (Treatment$_g$) and fixed effects for regions ($_g$ and $_{gt}$). Column (B) uses the municipality-level treatment variable and fixed effects for municipalities. Columns (C)–(D) are analogous to columns (A)–(B), but regressions include dummies for eleven municipality population groups fully interacted with dummies for each exam cohort and SES group. Columns (E)–(F) include dummies for full interactions between municipality population groups, municipality proximity groups, exam cohorts, and SES groups. See the text and Appendix Table A6 for details on the municipality groups.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A4. Reform effects on major choice

	(A)	(B)	(C)	(D)	(E)	(F)
	Dep. variable: Science & engineering major		Dep. variable: Business & economics major		Dep. variable: Law & social science major	
	Region-level	Muni-level	Region-level	Muni-level	Region-level	Muni-level
All students	0.002	0.003	0.003	−0.000	−0.003	−0.000
	(0.006)	(0.002)	(0.006)	(0.003)	(0.003)	(0.001)
Top income quartile	−0.013	−0.006	0.009	0.006	−0.013*	−0.004
	(0.013)	(0.006)	(0.011)	(0.005)	(0.007)	(0.003)
Bottom income quartile	0.006	0.009***	0.002	−0.004	0.000	0.001
	(0.007)	(0.003)	(0.007)	(0.003)	(0.004)	(0.002)
College educated mother	−0.011	−0.005	0.007	0.004	−0.006	−0.002
	(0.012)	(0.005)	(0.010)	(0.004)	(0.006)	(0.003)
Primary educated mother	0.005	0.007***	0.002	−0.003	0.001	0.001
	(0.006)	(0.002)	(0.006)	(0.003)	(0.004)	(0.002)
High ranked high school	−0.006	−0.003	0.004	0.003	−0.006	−0.002
	(0.011)	(0.005)	(0.010)	(0.005)	(0.005)	(0.003)
Low ranked high school	−0.001	0.004	0.006	−0.002	−0.000	0.001
	(0.007)	(0.003)	(0.007)	(0.003)	(0.004)	(0.002)

Notes: This table displays θ coefficients from separate regressions (7). The dependent variables are indicators for enrolling in the major listed in the column header, where majors are defined by the Ministry of Education's nine program areas. The omitted category includes health, education, fine arts, and veterinary sciences majors. The mean pre-reform enrollment rate in each major is: science & engineering (34.8%); business & economics (27.0%); law & social sciences (13.7%); and omitted category (24.6%).

The sample includes students in Column (B) of Table 1. The first row shows estimates for all students; other rows report estimates by SES group.

Columns (A), (C), and (E) report estimates using the region-level treatment variable, Treatment_g, and region dummies, $_g$. Columns (B), (D), and (F) use the municipality-level treatment variable and municipality dummies.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5. Reform effects on exam retaking and overall college enrollment

	(A)	(B)	(C)	(D)
	Dep. variable: Retook exam within one year		Dep. variable: Enrolled in any college	
	Region-level	Muni-level	Region-level	Muni-level
All students	−0.010	−0.006	−0.003	−0.000
	(0.013)	(0.005)	(0.009)	(0.004)
Top income quartile	−0.013	−0.005	−0.022	−0.008
	(0.013)	(0.006)	(0.015)	(0.008)
Bottom income quartile	−0.005	−0.004	0.001	0.002
	(0.011)	(0.004)	(0.007)	(0.003)
College educated mother	−0.010	−0.004	−0.001	−0.001
	(0.014)	(0.006)	(0.012)	(0.007)
Primary educated mother	−0.009	−0.006	0.003	0.002
	(0.011)	(0.004)	(0.008)	(0.004)
High ranked high school	0.004	0.001	−0.008	−0.007
	(0.014)	(0.007)	(0.012)	(0.007)
Low ranked high school	-0.017^{**}	-0.008^{***}	0.004	0.005
	(0.007)	(0.003)	(0.008)	(0.004)

Notes: This table displays θ coefficients from separate regressions (7). The dependent variable in columns (A)–(B) is an indicator equal to one if the student retook the national exam within one year. The dependent variable in columns (C)–(D) is an indictor for enrolling in any college in the Ministry of Education records.

The sample for columns (C)–(D) includes all students in the 1998–2001 exam cohorts, i.e., those in column (A) of Table 1. The sample for columns (A)–(B) includes only the 1998–2000 exam cohorts because I do not observe exam retaking beyond 2001. The first row shows estimates for all students; other rows report estimates by SES group.

Columns (A) and (C) report estimates using the region-level treatment variable, Treatment_g, and region dummies, $_g$. Columns (B) and (D) use the municipality-level treatment variable and municipality dummies.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A6. Municipality population and proximity groups for robustness tests

	(A) Exam takers per cohort	(B) Km from public uni	(C) Municipalities
1	26,831–79,592	0	Bogota, *Cali*, Medellin
2	5,054–17,045	0	Barranquilla, *Bucaramanga*, Cartagena, *Cucuta*, *Ibague*, *Manizales*, Monteria, *Neiva*, *Pasto*, *Pereira*, *Popayan*, Santa Marta, *Villavicencio*
3	1,633–4,378	0	*Armenia*, *Buenaventura*, *Duitama*, *Florencia*, *Girardot*, *Palmira*, Quibdo, Riohacha, *Sincelejo*, *Sogamoso*, *Tunja*, *Valledupar*
4	1,633–4,378	6–14	Bello, *Dos Quebradas*, Envigado, *Floridablanca*, Itagui, Soledad
5	1,633–4,378	20–32	Cerete, *Soacha*
6	1,633–4,378	32–40	*Buga*
7	1,633–4,378	53–85	*Barrancabermeja*, *Tulua*
8	690–1,504	0	*Aguachica*, Arauca, *Chiquinquira*, *Fusagasuga*, *Ocana*, *Pamplona*
9	690–1,504	6–14	*Calarca*, Copacabana
10	690–1,504	14–20	Caldas, *Corozal*, *Jamundi*, *Piedecuesta*, *Santa Rosa de Cabal*, *Yumbo*
11	690–1,504	20–32	Apartado, Baranoa, *Cartago*, *Espinal*, *Florida*, *Giron*, Rionegro
12	690–1,504	40–53	Cienaga, Fundacion, Lorica, *Madrid*, *Mosquera*, Sabanalarga, Sahagun, *Santander de Quilichao*, *Zipaquira*
13	690–1,504	53–85	*Chia*, *Facatativa*, *Ipiales*, *La Dorada*, Magangue, Maicao, *Pitalito*, *San Gil*, *Yopal*
14	690–1,504	141–231	*Tumaco*
15–23	282–660	9 groups	0km (*1T*, 5C), 6–14km (*5T*, 3C), 14–20km (*4T*, 7C), 20–32km (*9T*, 4C), 32–40km (*6T*, 7C), 40–53km (*13T*, 6C), 53–85km (*15T*, 10C), 86–134km (*2T*, 5C), 274–719km (*0T*, 1C)
24–32	174–272	9 groups	0km (*0T*, 2C), 6–14km (*3T*, 1C), 14–20km (*4T*, 1C), 20–32km (*6T*, 7C), 32–40km (*8T*, 3C), 40–53km (*19T*, 7C), 53–85km (*15T*, 5C), 86–134km (*5T*, 3C), 141–231km (*2T*, 1C),
33–40	85–172	8 groups	6–14km (*8T*, 3C), 14–20km (*16T*, 4C), 20–32km (*34T*, 11C), 32–40km (*19T*, 8C), 40–53km (*39T*, 13C), 53–85km (*35T*, 14C), 86–134km (*6T*, 3C), 141–231km (*0T*, 1C)
41–49	54–84	9 groups	6–14km (*4T*, 0C), 14–20km (*7T*, 2C), 20–32km (*14T*, 10C), 32–40km (*16T*, 6C), 40–53km (*13T*, 5C), 53–85km (*33T*, 16C), 86–134km (*6T*, 6C), 141–231km (*1T*, 0C)274–719km (*1T*, 2C),
50–58	26–54	9 groups	6–14km (*8T*, 3C), 14–20km (*9T*, 3C), 20–32km (*23T*, 6C), 32–40km (*27T*, 3C), 40–53km (*20T*, 11C), 53–85km (*49T*, 18C), 86–134km (*9T*, 8C), 141–231km (*6T*, 0C), 274–719km (*0T*, 1C)
59–67	12–26	9 groups	6–14km (*12T*, 1C), 14–20km (*13T*, 0C), 20–32km (*24T*, 2C), 32–40km (*16T*, 3C), 40–53km (*25T*, 5C), 53–85km (*36T*, 10C), 86–134km (*8T*, 6C), 141–231km (*6T*, 1C), 274–719km (*1T*, 1C),
68–76	2–11	9 groups	6–14km (*4T*, 0C), 14–20km (*2T*, 1C), 20–32km (*6T*, 1C), 32–40km (*3T*, 0C), 40–53km (*6T*, 1C), 53–85km (*22T*, 1C), 86–134km (*6T*, 4C), 141–231km (*1T*, 2C), 274–719km (*1T*, 0C)

Notes: Column (A) uses Ward's method for hierarchical clustering to divide municipalities into ten groups by the number of national exam takers per cohort. I create a separate eleventh group for Bogotá, Cali, and Medellín.

Column (B) uses Ward's method to divide municipalities into ten groups based on distance (in kilometers) from the nearest public university. Distance is defined as the crow flies using the capital city of each municipality.

Regressions in columns (C)–(D) of Table 5 include fully interacted dummies for population group/exam cohort/SES. Columns (E)–(F) in Table 5 include dummies for population group/proximity group/exam cohort/SES.

Italicized municipalities are in treated regions (T); non-italicized municipalities are in control regions (C).

Table A7. Reform effects on one year college persistence

	(A)	(B)	(C)	(D)	(E)	(F)
	Dependent variable: One year college persistence					
	Benchmark specification (7)		Within population groups		Within population × proximity groups	
	Region-level	Muni-level	Region-level	Muni-level	Region-level	Muni-level
All students	-0.0142^{***}	-0.0096^{***}	-0.0092	-0.0084^{**}	-0.0073	-0.0076^{*}
	(0.0051)	(0.0028)	(0.0071)	(0.0036)	(0.0074)	(0.0042)
Top income quartile	-0.0175^{*}	-0.0120^{***}	-0.0069	-0.0088	-0.0042	-0.0089
	(0.0092)	(0.0042)	(0.0142)	(0.0062)	(0.0147)	(0.0070)
Bottom income quartile	-0.0205^{***}	-0.0126^{***}	-0.0181^{**}	-0.0118^{***}	-0.0150^{**}	-0.0103^{**}
	(0.0056)	(0.0032)	(0.0067)	(0.0037)	(0.0068)	(0.0038)
College educated mother	-0.0158^{*}	-0.0098^{**}	-0.0067	-0.0066	-0.0061	-0.0061
	(0.0082)	(0.0042)	(0.0119)	(0.0055)	(0.0121)	(0.0064)
Primary educated mother	-0.0134^{**}	-0.0098^{***}	-0.0125^{*}	-0.0103^{***}	-0.0108	-0.0098^{***}
	(0.0051)	(0.0027)	(0.0069)	(0.0034)	(0.0067)	(0.0035)
High ranked high school	-0.0153^{**}	-0.0101^{**}	-0.0084	-0.0081	-0.0064	-0.0073
	(0.0068)	(0.0039)	(0.0128)	(0.0061)	(0.0140)	(0.0077)
Low ranked high school	-0.0136^{*}	-0.0084^{**}	-0.0119	-0.0087^{**}	-0.0110	-0.0084^{**}
	(0.0072)	(0.0031)	(0.0081)	(0.0037)	(0.0079)	(0.0040)

This table displays θ coefficients from separate regressions (7). The dependent variable is an indicator equal to one if the student was still in college one year after enrolling. The sample includes students in Column (B) of Table 1. The first row shows estimates for all students; other rows report estimates by SES group.

Columns (A) reports estimates of equation (7) using the region-level treatment variable, Treatment_g, and region dummies, $_g$. Column (B) uses the municipality-level treatment variable and municipality dummies. Columns (C)–(D) are analogous to columns (A)–(B), but regressions include dummies for eleven municipality population groups fully interacted with dummies for exam cohort. Columns (E)–(F) include dummies for full interactions between municipality population groups, municipality proximity groups, and exam cohort. See the text and Appendix Table A6 for details on these municipality groups.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A8. Reform effects on college graduation

	(A)	(B)	(C)	(D)	(E)	(F)
	Dependent variable: College graduation					
	Benchmark specification (7)		Within population groups		Within population × proximity groups	
	Region-level	Muni-level	Region-level	Muni-level	Region-level	Muni-level
All students	-0.0153^{*}	-0.0082^{**}	-0.0159	-0.0089^{**}	-0.0147	-0.0087^{*}
	(0.0076)	(0.0039)	(0.0097)	(0.0043)	(0.0099)	(0.0051)
Top income quartile	-0.0270^{***}	-0.0108^{**}	-0.0230^{*}	-0.0098	-0.0221^{*}	-0.0105
	(0.0094)	(0.0051)	(0.0128)	(0.0064)	(0.0130)	(0.0073)
Bottom income quartile	-0.0100	-0.0077	-0.0158^{*}	-0.0104^{**}	-0.0135	-0.0088^{**}
	(0.0085)	(0.0046)	(0.0087)	(0.0040)	(0.0084)	(0.0040)
College educated mother	-0.0221^{**}	-0.0085	-0.0188	-0.0069	-0.0200	-0.0081
	(0.0100)	(0.0057)	(0.0131)	(0.0067)	(0.0139)	(0.0079)
Primary educated mother	-0.0065	-0.0065	-0.0103	-0.0088^{**}	-0.0089	-0.0072^{*}
	(0.0075)	(0.0040)	(0.0075)	(0.0034)	(0.0078)	(0.0042)
High ranked high school	-0.0193^{**}	-0.0106^{**}	-0.0214	-0.0111	-0.0212	-0.0119
	(0.0074)	(0.0045)	(0.0160)	(0.0080)	(0.0179)	(0.0106)
Low ranked high school	-0.0100	-0.0040	-0.0131	-0.0062	-0.0111	-0.0049
	(0.0106)	(0.0045)	(0.0104)	(0.0039)	(0.0105)	(0.0040)

This table displays θ coefficients from separate regressions (7). The dependent variable is an indicator for college graduation. The sample includes students in Column (B) of Table 1. The first row shows estimates for all students; other rows report estimates by SES group.

Columns (A) reports estimates of equation (7) using the region-level treatment variable, Treatment_g, and region dummies, $_g$. Column (B) uses the municipality-level treatment variable and municipality dummies. Columns (C)–(D) are analogous to columns (A)–(B), but regressions include dummies for eleven municipality population groups fully interacted with dummies for exam cohort. Columns (E)–(F) include dummies for full interactions between municipality population groups, municipality proximity groups, and exam cohort. See the text and Appendix Table A6 for details on these municipality groups.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A9. Reform effects on log daily earnings

	(A)	(B)	(C)	(D)	(E)	(F)
	Dependent variable: Log daily earnings					
	Benchmark specification (7)		Within population groups		Within population × proximity groups	
	Region-level	Muni-level	Region-level	Muni-level	Region-level	Muni-level
All students	-0.0128^{**}	-0.0075^{**}	-0.0133^{*}	-0.0077^{**}	-0.0175^{***}	-0.0112^{***}
	(0.0053)	(0.0027)	(0.0066)	(0.0036)	(0.0062)	(0.0035)
Top income quartile	-0.0178^{*}	-0.0081	-0.0253^{*}	-0.0100	-0.0340^{***}	-0.0126
	(0.0092)	(0.0055)	(0.0126)	(0.0079)	(0.0119)	(0.0085)
Bottom income quartile	-0.0133^{***}	-0.0077^{**}	-0.0154^{***}	-0.0085^{**}	-0.0165^{**}	-0.0101^{**}
	(0.0047)	(0.0032)	(0.0049)	(0.0036)	(0.0064)	(0.0038)
College educated mother	-0.0111	-0.0038	-0.0140	-0.0006	-0.0277^{*}	-0.0074
	(0.0108)	(0.0059)	(0.0141)	(0.0074)	(0.0138)	(0.0081)
Primary educated mother	-0.0098^{**}	-0.0075^{***}	-0.0112^{***}	-0.0086^{***}	-0.0097^{**}	-0.0088^{**}
	(0.0046)	(0.0025)	(0.0040)	(0.0028)	(0.0045)	(0.0033)
High ranked high school	-0.0060	-0.0062	-0.0064	-0.0054	-0.0126	-0.0071
	(0.0088)	(0.0050)	(0.0132)	(0.0067)	(0.0149)	(0.0094)
Low ranked high school	-0.0125^{**}	-0.0046	-0.0169^{***}	-0.0067^{**}	-0.0183^{***}	-0.0074^{**}
	(0.0055)	(0.0029)	(0.0052)	(0.0029)	(0.0056)	(0.0032)

This table displays θ coefficients from separate regressions (7). The dependent variable is log daily earnings measured 10–11 years after the admission exam. The sample includes the subset of students from Column (B) of Table 1 who have formal sector earnings. The first row shows estimates for all students; other rows report estimates by SES group.

Columns (A) reports estimates of equation (7) using the region-level treatment variable, Treatment_g, and region dummies, $\ _g$. Column (B) uses the municipality-level treatment variable and municipality dummies. Columns (C)–(D) are analogous to columns (A)–(B), but regressions include dummies for eleven municipality population groups fully interacted with dummies for exam cohort. Columns (E)–(F) include dummies for full interactions between municipality population groups, municipality proximity groups, and exam cohort. See the text and Appendix Table A6 for details on these municipality groups.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A10. Reform effects on different earning measures

	(A)	(B)	(C)	(D)	(E)	(F)
	Dep. variable: Log daily earnings 10–11 years later		Dep. variable: Log daily earnings in 2011–2012		Dep. variable: Log annual earnings 10–11 years later	
	Region-level	Muni-level	Region-level	Muni-level	Region-level	Muni-level
All students	-0.015^{***}	-0.009^{***}	-0.012	-0.009^{**}	-0.015	-0.012^{*}
	(0.005)	(0.003)	(0.009)	(0.003)	(0.013)	(0.006)
Top income quartile	-0.021^{**}	-0.009	-0.022^{*}	-0.012^{***}	-0.022	-0.004
	(0.010)	(0.006)	(0.011)	(0.005)	(0.020)	(0.011)
Bottom income quartile	-0.015^{***}	-0.009^{**}	-0.008	-0.006	-0.019	-0.023^{**}
	(0.005)	(0.004)	(0.011)	(0.005)	(0.017)	(0.010)
College educated mother	-0.011	-0.003	-0.012	-0.006	0.006	0.008
	(0.012)	(0.006)	(0.013)	(0.006)	(0.025)	(0.013)
Primary educated mother	-0.012^{**}	-0.009^{***}	-0.007	-0.008^{**}	-0.008	-0.018^{**}
	(0.005)	(0.003)	(0.010)	(0.004)	(0.014)	(0.007)
High ranked high school	-0.008	-0.006	-0.011	-0.010	-0.004	-0.003
	(0.009)	(0.005)	(0.014)	(0.006)	(0.017)	(0.010)
Low ranked high school	-0.012^{**}	-0.005^{*}	-0.010	-0.005	-0.006	-0.011^{*}
	(0.005)	(0.003)	(0.009)	(0.004)	(0.011)	(0.006)

Notes: This table displays θ coefficients from separate regressions (7) with three dependent variables: log daily earnings 10–11 years after the admission exam (columns (A)–(B)); log daily earnings in 2011–2012 (columns (C)–(D)); and log *annual* earnings 10–11 years after the admission exam (columns (E)–(F)). The sample includes students in Column (B) of Table 1 who have a formal sector earnings for *each* of the three variables. The first row shows estimates for all students; other rows report estimates by SES group.

Columns (A), (C), and (E) report estimates using the region-level treatment variable, Treatment_g, and region dummies, $_g$. Columns (B), (D), and (F) use the municipality-level treatment variable and municipality dummies.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A11. Reform effects on formal employment

	(A)	(B)	(C)	(D)	(E)	(F)
	Dependent variable: Appears in formal earnings records					
	Benchmark specification (7)		Within population groups		Within population × proximity groups	
	Region-level	Muni-level	Region-level	Muni-level	Region-level	Muni-level
All students	-0.0102	-0.0079^{**}	-0.0143	-0.0097^{**}	-0.0159	-0.0114^{***}
	(0.0102)	(0.0039)	(0.0112)	(0.0036)	(0.0122)	(0.0036)
Top income quartile	-0.0200	-0.0139^{**}	-0.0231	-0.0153^{**}	-0.0258	-0.0179^{***}
	(0.0165)	(0.0059)	(0.0196)	(0.0063)	(0.0214)	(0.0064)
Bottom income quartile	-0.0040	-0.0064	-0.0130^{*}	-0.0102^{**}	-0.0118	-0.0096^{**}
	(0.0088)	(0.0046)	(0.0069)	(0.0038)	(0.0079)	(0.0043)
College educated mother	-0.0171	-0.0138^{*}	-0.0252	-0.0179^{**}	-0.0273	-0.0205^{***}
	(0.0174)	(0.0068)	(0.0209)	(0.0068)	(0.0238)	(0.0074)
Primary educated mother	-0.0017	-0.0039	-0.0037	-0.0047	-0.0043	-0.0058
	(0.0073)	(0.0037)	(0.0079)	(0.0037)	(0.0085)	(0.0038)
High ranked high school	-0.0143	-0.0114	-0.0244	-0.0163^{**}	-0.0319	-0.0242^{***}
	(0.0179)	(0.0068)	(0.0226)	(0.0066)	(0.0283)	(0.0073)
Low ranked high school	-0.0113	-0.0073^{**}	-0.0123	-0.0077^{**}	-0.0084	-0.0056^{*}
	(0.0077)	(0.0035)	(0.0082)	(0.0030)	(0.0077)	(0.0028)

This table displays θ coefficients from separate regressions (7). The dependent variable is an indicator for appearing in my earnings records. The sample includes students in Column (B) of Table 1. The first row shows estimates for all students; other rows report estimates by SES group.

Columns (A) reports estimates of equation (7) using the region-level treatment variable, Treatment_g, and region dummies, $_g$. Column (B) uses the municipality-level treatment variable and municipality dummies. Columns (C)–(D) are analogous to columns (A)–(B), but regressions include dummies for eleven municipality population groups fully interacted with dummies for exam cohort. Columns (E)–(F) include dummies for full interactions between municipality population groups, municipality proximity groups, and exam cohort. See the text and Appendix Table A6 for details on these municipality groups.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A12. Clustered and wild t bootstrap p values for reform effects on SES college quality gaps

	(A)	(B)
Dependent variable: College quality, Q_c		
	Benchmark specification (8)	
	Region-level	Muni-level
Family income gap	−2.402	−1.561
Top Q − Bottom Q	(0.000)	(0.000)
	[0.004]	[0.004]
Mother's education income gap	−1.409	−0.803
College − Primary	(0.007)	(0.015)
	[0.048]	[0.052]
High school rank gap	−1.476	−1.308
High − Low	(0.119)	(0.011)
	[0.180]	[0.020]

Notes: This table presents regression coefficients identical to columns (A)–(B) in Table 5. Parentheses contain p values from standard errors clustered at the region level, as in Table 5. Brackets contain p values from a wild t bootstrap with 500 replications that imposes the null hypothesis, as recommended in Cameron et al. (2008).

Table A13. Clustered and wild t bootstrap p values for reform effects on college persistence, graduation, and earnings

	(A)	(B)	(C)	(D)	(E)	(F)
	Dep. variable: One year persistence		Dep. variable: College graduation		Dep. variable: Log daily earnings	
	Region-level	Muni-level	Region-level	Muni-level	Region-level	Muni-level
All students	−0.014	−0.010	−0.015	−0.008	−0.013	−0.007
	(0.009)	(0.002)	(0.051)	(0.043)	(0.021)	(0.010)
	[0.052]	[0.012]	[0.064]	[0.040]	[0.056]	[0.020]
Top income quartile	−0.017	−0.012	−0.027	−0.011	−0.018	−0.008
	(0.065)	(0.007)	(0.007)	(0.042)	(0.063)	(0.150)
	[0.140]	[0.016]	[0.020]	[0.056]	[0.088]	[0.208]
Bottom income quartile	−0.021	−0.013	−0.010	−0.008	−0.013	−0.008
	(0.001)	(0.000)	(0.251)	(0.106)	(0.008)	(0.020)
	[0.020]	[0.036]	[0.307]	[0.160]	[0.024]	[0.072]
College educated mother	−0.016	−0.010	−0.022	−0.008	−0.011	−0.004
	(0.065)	(0.025)	(0.034)	(0.143)	(0.310)	(0.523)
	[0.196]	[0.028]	[0.096]	[0.236]	[0.331]	[0.555]
Primary educated mother	−0.013	−0.010	−0.007	−0.007	−0.010	−0.007
	(0.013)	(0.001)	(0.393)	(0.114)	(0.041)	(0.005)
	[0.080]	[0.024]	[0.371]	[0.136]	[0.068]	[0.004]
High ranked high school	−0.015	−0.010	−0.019	−0.011	−0.006	−0.006
	(0.032)	(0.014)	(0.013)	(0.025)	(0.504)	(0.220)
	[0.112]	[0.044]	[0.028]	[0.028]	[0.563]	[0.287]
Low ranked high school	−0.014	−0.008	−0.010	−0.004	−0.013	−0.005
	(0.068)	(0.011)	(0.354)	(0.387)	(0.029)	(0.114)
	[0.076]	[0.044]	[0.395]	[0.451]	[0.032]	[0.112]

Notes: This table presents regression coefficients identical to those in Table 6. Parentheses contain p values from standard errors clustered at the region level, as in Table 6. Brackets contain p values from a wild t bootstrap with 500 replications that imposes the null hypothesis, as recommended in Cameron et al. (2008).

Table A14. Reform effects on SES college quality gaps excluding certain regions
Dependent variable: College quality, Q_c

	(A)	(B)	(C)
	All regions	Exclude small regions	Exclude mixed admit
Family income gap	-2.40^{***}	-2.47^{***}	-2.01^{**}
Top Q − Bottom Q	(0.54)	(0.54)	(0.74)
Mother's education gap	-1.41^{***}	-1.44^{***}	-0.90
College − Primary	(0.49)	(0.50)	(0.65)
High school rank gap	-1.48	-1.54	-0.65
High − Low	(0.92)	(0.94)	(1.07)

Notes: This table displays θ^q coefficients from separate regressions (8). The sample includes students in Column (B) of Table 1. Rows are defined by the three SES measures from Table 1.

Columns (A) reports estimates from equation (8) using the region-level treatment variable (Treatment_g) and fixed effects for regions ($_g$ and $_{gt}$). Column (B) is analogous to column (A), but it omits the seven regions with no public universities (Casanare, Guainía, Guaviare, Putumayo, San Andrés, Vaupés, and Vichada). Column (C) is analogous to column (B), but it also omits the three regions with mixed admission methods (Bogotá, Caldas, and Valle del Cauca). See Appendix Table A2 for details on the regions.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A15. Reform effects on college graduation and earnings excluding certain regions

	(A)	(B)	(C)	(D)	(E)	(F)
	Dep. variable: College graduation			Dep. variable: Log daily earnings		
	All regions	Exclude small regions	Exclude mixed admit	All regions	Exclude small regions	Exclude mixed admit
All students	-0.015^{*}	-0.015^{*}	-0.015	-0.013^{**}	-0.014^{**}	-0.016^{*}
	(0.008)	(0.008)	(0.013)	(0.005)	(0.005)	(0.009)
Top income quartile	-0.027^{***}	-0.027^{***}	-0.028^{*}	-0.018^{*}	-0.020^{**}	-0.023
	(0.009)	(0.009)	(0.015)	(0.009)	(0.009)	(0.016)
Bottom income quartile	-0.010	-0.010	-0.012	-0.013^{***}	-0.015^{***}	-0.015^{**}
	(0.009)	(0.009)	(0.013)	(0.005)	(0.004)	(0.006)
College educated mother	-0.022^{**}	-0.022^{**}	-0.019	-0.011	-0.013	-0.012
	(0.010)	(0.010)	(0.017)	(0.011)	(0.011)	(0.020)
Primary educated mother	-0.007	-0.006	-0.004	-0.010^{**}	-0.012^{***}	-0.013^{**}
	(0.008)	(0.008)	(0.011)	(0.005)	(0.004)	(0.005)
High ranked high school	-0.019^{**}	-0.019^{**}	-0.023	-0.006	-0.008	-0.011
	(0.007)	(0.007)	(0.017)	(0.009)	(0.009)	(0.015)
Low ranked high school	-0.010	-0.009	-0.012	-0.013^{**}	-0.014^{**}	-0.017^{**}
	(0.011)	(0.011)	(0.013)	(0.005)	(0.005)	(0.006)

Notes: This table displays θ coefficients from separate regressions (7) with two dependent variables: an indicator for college graduation (columns (A)–(C)); and log daily earnings 10–11 years after the admission exam (columns (D)–(F)). The sample in columns (A)–(C) includes students in Column (B) of Table 1; columns (D)–(F) include only students with formal sector earnings. The first row shows estimates for all students; other rows report estimates by SES group.

All columns report estimates using the region-level treatment variable, Treatment_g, and region dummies, $_g$. Columns (A) and (D) include all regions. Columns (B) and (E) omit the seven regions with no public universities (Casanare, Guainía, Guaviare, Putumayo, San Andrés, Vaupés, and Vichada). Columns (C) and (F) omit these seven regions plus the three regions with mixed admission methods (Bogotá, Caldas, and Valle del Cauca). See Appendix Table A2 for details on the regions.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A16. Reform effects on the correlation of peer characteristics and college quality, Q_c

	(A)	(B)	(C)	(D)
	Dependent variable (in st. deviation units)			
	% from top income quartile	% with college edu. mom	% from high ranked high school	Mean s. science exam score
College quality, Q_c	-0.030	0.005	0.016	-0.003
	(0.033)	(0.053)	(0.025)	(0.030)
N	1,167	1,167	1,167	1,167
R^2	0.541	0.547	0.655	0.739
Mean pre-reform corr.	0.677	0.696	0.756	0.867

Notes: This table displays estimates of reform effects on the correlation of peer characteristics and college quality, Q_c. Regressions are at the college/exam cohort level. The dependent variable is the mean peer characteristic (within a given college and exam cohort) listed in the column header. The table displays the coefficient on the interaction between Q_c, the region-level treatment variable, Treated_g, and a dummy for the post-reform cohorts, Post_t. Regressions also include region-cohort dummies, interactions between Q_c and region dummies, and interactions between Q_c and exam cohort dummies.

In all regressions, variables are normalized to standard deviation one within treated/control regions and exam cohorts so that coefficients can be interpreted as correlation coefficients.

The bottom row shows the correlation between the dependent variable and college quality, Q_c, for the 1998–1999 cohorts. Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A17. IV estimates of returns to college quality, $_z$ (municipality-level treatment)

Panel A. Average returns by SES (standard 2SLS estimation)

	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
		First stage effects, θ_z			Returns to college quality, $_z$			
	N	Coef	SE	F stat	Coef	SE	# coefs	St. dev
Top income quartile	84,615	-0.117	(0.020)	34.6	0.070	(0.050)	1	.
Bottom income quartile	83,587	0.032	(0.039)	0.7	-0.240	(0.258)	1	.
College educated mother	85,604	-0.075	(0.027)	7.6	0.051	(0.084)	1	.
Primary educated mother	138,655	0.009	(0.036)	0.1	-0.845	(3.211)	1	.
High ranked high school	118,394	-0.087	(0.031)	8.0	0.071	(0.072)	1	.
Low ranked high school	102,874	0.018	(0.030)	0.4	-0.258	(0.447)	1	.
All students	340,623	-0.035	(0.029)	1.4	0.215	(0.234)	1	.

Panel B. Across- and within-SES returns (IV random forest estimation)

		First stage effects, θ_z			Returns to college quality, $_z$			
	N	Mean coef	Mean SE	N with $F > 10$	Mean coef	Mean SE	# coefs	St. dev
Top income quartile	35,287	-0.088	0.064	3,545	0.094	0.067	487	0.036
Bottom income quartile	31,817	0.021	0.048	317	-0.088	0.104	48	0.096
College educated mother	36,164	-0.061	0.059	2,709	0.090	0.064	283	0.040
Primary educated mother	50,940	0.012	0.047	385	-0.062	0.097	94	0.108
High ranked high school	48,883	-0.066	0.061	3,537	0.094	0.067	476	0.036
Low ranked high school	39,159	0.021	0.047	366	-0.095	0.098	59	0.089
All students	135,200	-0.011	0.052	4,523	0.059	0.070	676	0.082

Notes: Panel A shows estimates from equations (9)–(10) using municipality-level treatment and covariate groups z defined by the SES group in the first column. Column (A) shows the sample size for each regression, which includes students in column (B) of Table 1 with formal sector earnings. Columns (B)–(C) report estimates of θ_z and its standard error from equation (9). Column (D) shows F statistics from tests of $\theta_z = 0$. Columns (E)–(F) report estimates of $_z$ and its standard error from equation (10). Column (G) shows that there is one coefficient per SES group (in contrast to Panel B). Standard errors in columns (C) and (F) are clustered at the region level.

Panel B summarizes Athey et al. (2017) random forest estimates from equations (9)–(10) (see Appendix C.5). The estimation uses a 50 percent training sample, and this table presents statistics from the validation sample. Column (A) shows the number of students in the validation sample, which excludes students with missing covariates, Z_i. Columns (B)–(C) show the mean of the θ_z estimates and the average standard error. Column (D) shows the number of students in the validation sample with first stage F statistics above ten; columns (E)–(H) present statistics only from this subsample. Columns (E)–(F) show the mean of the $_z$ estimates and the average standard error. Column (G) shows the number of unique $_z$ parameters. Column (H) reports the standard deviation of the $_z$ estimates.

In all regressions, I divide Q_c by ten so that one unit equals ten percentile points in the distribution of colleges.

Table A18. Reform effects on the correlation of returns to college quality, $_z$, and admission scores (municipality-level treatment)

	(A)	(B)	(C)	(D)
	Dependent variable: Return to college quality, $_z$			
	All students		Bottom income quartile	
Exam subject	Pre-reform correlation	Reform effect	Pre-reform correlation	Reform effect
Math	0.394***	-0.289***	0.296***	-0.411**
	(0.035)	(0.045)	(0.115)	(0.175)
Language	0.336***	-0.082*	0.302***	-0.214
	(0.036)	(0.047)	(0.116)	(0.174)
Biology	0.363***	-0.122***	0.294**	-0.204
	(0.036)	(0.046)	(0.123)	(0.175)
Chemistry	0.370***	-0.094**	0.365***	-0.169
	(0.036)	(0.046)	(0.118)	(0.173)
Physics	0.391***	-0.228***	0.185*	-0.229
	(0.034)	(0.044)	(0.112)	(0.173)
Social science	0.356***	-0.092**	0.377***	-0.211
	(0.036)	(0.047)	(0.120)	(0.177)
N	1,763	4,523	166	317

Notes: The sample includes students from column (D) in Table A17 (Panel B). Column (A) reports coefficients from separate regressions of random forest returns, $_z$, on admission scores for each subject using the 1998–1999 cohorts. Column (B) shows the change in these coefficients between the 2000–2001 and 1998–1999 cohorts. Columns (C)–(D) are analogous, but regressions include only bottom income quartile students. All variables are normalized to standard deviation one within the pre- and post-reform cohorts so coefficients can be interpreted as bivariate correlations.

Estimates of $_z$ use the municipality-level treatment variable; Table 8 replicates this table with region-level estimates of $_z$. Parentheses contain robust standard errors adjusted for estimation error. For this I form a diagonal matrix where each element is the squared residual from regressing $_z$ on the admission score plus the squared standard error of $_z$. I then use this diagonal error matrix to compute robust standard errors.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B. Theoretical appendix

B.1. **Derivation of Proposition 1.** This section provides details on Proposition 1 in Section 1, which describes how reducing SES test score gaps in college admission exams can affect labor market outcomes.

I consider labor market outcomes related to both efficiency and equity. A natural measure of efficiency is average log earnings in the market, $\bar{w} = E[w_{ic}]$, which is given by:[43]

$$
\begin{aligned}
\bar{w} &= \bar{\ } + E[\ _iQ_c] \\
&= \bar{\ } + E[\ _iT_i] \\
&= \bar{\ } + \text{Cov}(\ _i, T_i) && \text{(B1)} \\
&= \bar{\ } + \frac{(\bar{\ }_1 - \bar{\ }_0)(\bar{T}_1\psi - \bar{T}_0\psi)}{4} + \sum_{x=0}^{1} \frac{\text{Cov}(\ _i, T_i|X_i = x)}{2} && \text{(B2)}
\end{aligned}
$$

where $\bar{\ } = E[\ _i]$ and $\bar{\ }_1 - \bar{\ }_0 = E[\ _i|X_i = 1] - E[\ _i|X_i = 0]$. Equations (B1) and (B2) are equivalent represents of mean earnings that are useful for the proposition below.

Policymakers may also be interested in how college admission exams affect earnings inequality. Average log earnings for SES group x, which I denote by $\bar{w}_x = E[w_{ic}|X_i = x]$, is given by:

$$
\begin{aligned}
\bar{w}_x &= \bar{\ }_x + E[\ _iQ_c|X_i = x] \\
&= \bar{\ }_x + E[\ _iT_i|X_i = x] \\
&= \bar{\ }_x + \bar{\ }_x\bar{T}_x\psi + \text{Cov}(\ _i, T_i|X_i = x).\psi && \text{(B3)}
\end{aligned}
$$

It follows that the SES earnings gap, $\bar{w}_1 - \bar{w}_0$ is

$$
\bar{w}_1 - \bar{w}_0 = \bar{\ }_1 - \bar{\ }_0 + \bar{\ }(\bar{T}_1\psi - \bar{T}_0\psi) + \text{Cov}(\ _i, T_i|X_i = 1) - \text{Cov}(\ _i, T_i|X_i = 0),\psi \quad \text{(B4)}
$$

where $\bar{\ } = E[\ _i]$.[44] Equations (B3) and (B4) capture the implications of admission tests for SES-specific earnings and the degree of earnings inequality in the market.

Equations (B1)–(B4) show that the implications of exam reform for earnings equity and efficiency depend on the characteristics of the return to college quality, $\ _i$, and on the exam's predictive power for these returns. To illustrate this, Proposition 1 describes the earnings effects of reducing the SES test score gap under three different assumptions on the distribution of $\ _i$.

[43] The last line follows from the covariance decomposition formula, $\text{Cov}(\ , T) = \text{Cov}\big(E[\ |X], E[T|X]\big) + E\big[\text{Cov}(\ , T|X)\big]$, and from the fact that $\bar{T}_1 - \bar{T}_0 = 2\bar{T}_1$.

[44] This follows from $\bar{T}_1 - \bar{T}_0 = 2\bar{T}_1$, and $\bar{\ } = \frac{1}{2}(\bar{\ }_1 + \bar{\ }_0)$.

Proposition 1. *Consider an exam reform that reduces the SES test score gap,* $\bar{T}_1 - \bar{T}_0$*:*

- *(A)* ***Constant returns.*** *If the return to college quality is positive and constant for all students (* $_i = \; > 0$*), mean earnings is unaffected and earnings inequality decreases.*
- *(B)* ***Complementarity.*** *If returns are larger for high SES students on average (* $\bar{}_1 > \bar{}_0$*), mean earnings decreases unless the exam becomes a better predictor of returns within SES groups, i.e.,* $\text{Cov}(\;_i, T_i | X_i = x)$ *increases.*
- *(C)* ***Mismatch.*** *Mean earnings for low SES students can* <u>*decrease*</u> *if either:*
  - *The mean return to college quality is negative for low SES students (* $\bar{}_0 < 0$*); or*
  - *The exam becomes a worse predictor of low SES returns, i.e.,* $\text{Cov}(\;_i, T_i | X_i = 0)$ *decreases.*

Below I provide details on parts (A)–(C) of Proposition 1.

(A) *Constant returns.* Consider first the simple case in which the return to college quality is positive and constant across individuals, i.e., $_i = \; > 0$ for all i. Since $_i$ is fixed we have $\text{Cov}(\;_i, T_i) = \text{Cov}(\;_i, T_i | X_i = x) = 0$ in equations (B1) and (B4). Average earnings reduce to $\bar{w} = \bar{}$, and the SES earnings gap becomes $\bar{w}_1 - \bar{w}_0 = \bar{}_1 - \bar{}_0 + \; (\bar{T}_1 - \bar{T}_0)$.

With a constant $_i$, reforms that decrease the test score gap, $\bar{T}_1 - \bar{T}_0$, reduce earnings inequality without changing average earnings. Intuitively, when there is no heterogeneity in the return to college quality, admission test design has no efficiency implications because all students get the same benefit from attending better colleges. Lowering the test score gap raises earnings for low SES students and reduces earnings for high SES students, but these effects are exactly offsetting.

(B) *Complementarity.* Suppose now that $_i$ is heterogenous and that the average return to college quality is larger for high SES students, i.e., $\bar{}_1 > \bar{}_0 > 0$. Heterogeneity in $_i$ means that tests that assign higher scores to students with larger returns college quality will raise average earnings.[45]

If high SES students have larger returns to college quality on average, then average earnings are increasing in the test score gap because $\bar{}_1 - \bar{}_0 > 0$ in equation (B2). Exam reforms that reduce test score gaps will lower average earnings unless the new test becomes more related to $_i$ *within* SES groups, i.e., unless $\text{Cov}(\;_i, T_i | X_i = x)$ increases. This covariance term depends on the types of questions in the exam, but it also depends on the within-SES variance in $_i$. If $_i$ is strongly related to SES

[45] For example, equation (B1) shows that average earnings are maximized by the test that maximizes $\text{Cov}(\;_i, T_i)$.

but varies little within SES groups, lowering test score gaps will necessarily reduce average earnings.[46]

Thus when there is a complementarity between SES and returns to college quality, there is an efficiency consideration in the design of admission tests. Admission tests that assign higher scores to low SES students can lower market-wide earnings.

(C) *Mismatch.* Part (B) of Proposition 1 describes a case in which reducing the SES test score gap can lower average earnings in the market. But there are conditions on β_i in which reducing test score gaps can also reduce earnings for low SES students.

One such condition is if the average return to college quality is negative for low SES students, $\bar{\beta}_0 < 0$. In this case, equation (B3) shows that increasing low SES test scores, $\bar{T}_0$, lowers average low SES earnings, $\bar{w}_0$, all else equal. If most low SES students have negative returns to college quality, then decreasing the test score gap shifts many low SES students into higher quality colleges where they are less likely to succeed. The case of $\bar{\beta}_0 < 0$ is often called the "mismatch hypothesis," which argues that some students may be better off attending lower-ranked schools if they are academically unprepared for top colleges (Arcidiacono and Lovenheim, 2016).

Even if $\bar{\beta}_0 > 0$, reforms that raise low SES students' test scores can reduce average low SES earnings if $\text{Cov}(\beta_i, T_i | X_i = 0)$ decreases (equation (B3)). A reduction in $\text{Cov}(\beta_i, T_i | X_i = 0)$ means that low SES students are assigned to the "wrong" colleges; students with higher returns to college quality are less likely to receive high test scores, while students with lower values of β_i are more likely to score well on the exam. This is not mismatch in the sense that students are worse off at better colleges, but students are not matched to colleges in a way that maximizes their average earnings.[47] In this sense, reforms that raise low SES test scores can lead to mismatch if the new test is a poor predictor of which students are likely to benefit the most at top colleges.

B.2. **Exam validity and returns to college quality.** Section 1 shows that if the goal of a college admission exam is to promote efficiency and equity in earnings, the test should be a good predictor of each student's return to college quality, β_i. In this sense, the ideal exam assigns high weight to abilities a_{ki} that are strongly correlated with β_i but weakly correlated with SES. It may not be easy to identify such abilities, especially when high SES students have easier access to test prep services that are likely to respond to any exam reform.

[46] $\text{Cov}(\beta_i, T_i | X_i = x)$ is bounded by the variance of β_i within SES groups since $\text{Cov}(\beta_i, T_i | X_i = x) \leq \sqrt{\text{Var}(\beta_i | X_i = x) \times \text{Var}(T_i | X_i = x)}$. Thus when $\text{Var}(\beta_i | X_i = x)$ is small, the primary effect of the student-college match on average earnings is $(\bar{\beta}_1 - \bar{\beta}_0)(\bar{T}_1 - \bar{T}_0)$ (equation (B2)).

[47] However, unless a substantial fraction of low SES students have negative returns, $\beta_i < 0$, reducing the test score gap is unlikely to lower average low SES earnings. This follows from the discussion in footnote 46; if $\bar{\beta}_0$ is large and positive and $\text{Var}(\beta_i | X_i = 0)$ is small, then increasing $\bar{T}_0$ raises $\bar{w}_0$ (equation (B3)).

Furthermore, $_i$ is the causal return to attending different colleges, which is a difficult to parameter to observe.

The testing agency may be able to get around these hurdles if it can measure other characteristics that are strongly correlated with $_i$. In this section, I compare the goal of measuring $_i$ to what testing agencies do in practice to evaluate their exams. Testing agencies typically measure an exam's validity by correlating test scores with college outcomes. Let Y_i be a measure of college success such as GPA or graduation, and let σ_Y^2 denote its variance. An exam's "raw validity," which I denote by ρ_{YT}, is the correlation coefficient between test scores and college outcomes:

$$\rho_{YT} = \frac{\text{Cov}(Y_i, T_i)}{\sigma_Y}.\psi$$

To relate validity to the model, let the linear projection of returns to college quality, $_i$, onto the the measure of college success, Y_i, be given by

$$_i = \pi\psi + \phi Y_i + \tilde{}_i,\psi$$

where $\phi\psi = \text{Cov}(_i, Y_i)/\sigma_Y^2$. Then from equation (B1), average earnings can be written as:

$$\begin{aligned}
\bar{w} &= \bar{} + \text{Cov}(\pi\psi + \phi Y_i + \tilde{}_i, T_i) \\
&= \bar{} + \phi\text{Cov}(Y_i, T_i) + \text{Cov}(\tilde{}_i, T_i) \\
&= \bar{} + \sigma\ \rho_{Y}\ \rho_{YT} + \text{Cov}(\tilde{}_i, T_i) \qquad \text{(B5)}
\end{aligned}$$

where ρ_{Y} is the correlation between Y_i and $_i$, and σ^2 is the variance of $_i$.

Equation (B5) shows that exam validity, ρ_{YT}, matters for average earnings only to the extent that the there is a strong correlation between the measure of college success, Y_i, and the return to college quality, $_i$. If ρ_{Y} is large and positive, then increasing exam validity raises average earnings. If ρ_{Y} is small, then maximizing exam validity has a limited effect on students' labor market outcomes; average earnings is primarily determined by the part of $_i$ that is unrelated to Y_i.

Equation (B5) relates exam validity to average earnings, but a similar idea carries over to earnings inequality. In addition to raw validity, $\rho_{T,Y}$, testing agencies commonly calculate correlations of Y_i and T_i for different SES groups, which is often called "differential validity." An analogous decomposition of equation (B3) shows that the relationship between differential validity and SES-specific earnings depends on the correlation of Y_i and $_i$ *within* SES groups. Thus maximizing validity within SES groups can guard against mismatch only to the extent that Y_i is strongly related to $_i$ within SES groups.

C. Empirical appendix

Table C1. Construction of analysis sample

	N
Total number of exam takers	2,100,424
Missing exam scores	(10,195)
Missing high school information	(126,966)
Fewer than five pre-reform obs. in gender/HS/mom ed cells	(319,003)
Full sample	1,644,260
College enrollees	612,949

Notes: See the text for descriptions of the sample restrictions in each row.

C.1. **Data, sample, and variable definitions.** This section describes the coverage and merging of my three main administrative datasets: college admission exam records, enrollment and graduation records, and earnings records. It also describes my analysis sample and definitions of key variables.

The first dataset includes records from the ICFES national standardized college entrance exam. The data include all students who took the exam between 1998–2001. My sample includes all exam takers with non-missing test scores and high school identifiers. In addition, I make a sample restriction that allows me to calculate a consistent measure of family income across cohorts. In the data, family income is grouped into ten bins based on multiples of the monthly minimum wage, but the distribution of these bins changes dramatically across cohorts due to variation in inflation and minimum wage policy. To get a stable measure of family income, I calculate predicted family income using an individual's gender, mother's education, and high school. Specifically, I define predicted income as the mean family income as fraction of the minimum wage within cells defined by gender, nine mother's education categories, and the roughly 7,500 high schools in my sample. I use only 1998–1999 cohorts to predict family income, and I drop any cells with fewer than five observations in the pre-reform cohorts. I then define income quartiles based on a student's percentile rank of predicted family income within their exam cohort.

Table C1 shows the effect of these restrictions on sample size. The full sample includes roughly 1.6 million exam takers. Table C2 shows that I find no evidence of differential selection into the sample in treatment and control areas. Most of my analyses are restricted to those who enrolled in college, as shown in the last row of Table C1. I find no evidence of reform effects on the probability of college enrollment (Appendix Table A5).

In addition to family income quartiles, I use two other SES measures computed from the ICFES records. First, I use two mother education groups: students whose mothers

Table C2. Balance test for missing data
Dependent variable: Appears in analysis sample

	(A)	(B)
	Region-level	Muni-level
All students	−0.006	−0.001
	(0.008)	(0.004)
N	1,962,923	1,962,923
Mean	0.838	0.838

Notes: This table displays θ coefficients from separate regressions (7). The dependent variable is an indicator equal to one if the student appears in the analysis sample (fifth row of Table C1). The sample includes all exam takers except those with missing exam scores and missing high school information (rows 2–3 of Table C1).

Column (A) reports estimates using the region-level treatment variable, Treatment_g, and region dummies, $_g$. Column (B) uses the municipality-level treatment variable and municipality dummies.

Parentheses contain standard errors clustered at the region level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

attended college (or more) and students whose mothers attended primary school or less. Second, I use groups defined by high school rank. The ICFES testing agency classifies high schools into seven categories based on their mean admission exam performance. The low rank group includes the bottom three categories, and the high rank group includes the top three categories. I use a high school's pre-reform rank and hold this definition fixed across cohorts.

The second dataset includes enrollment and graduation records from the Ministry of Education. The Ministry's records include almost all colleges in Colombia, although it omits a few schools due to their small size or inconsistent reporting. To describe the set of colleges that are included in the Ministry of Education records, I use another administrative dataset from a college exit exam called *Saber Pro* (formerly ECAES). This national exam is administered by the same agency that runs the ICFES college admission exam. The exit exam became a requirement for graduation from any higher education institution in 2009.

Column (A) in Table C3 depicts the 310 colleges that have any exit exam takers in these administrative records in 2009–2011. These colleges are categorized into the Ministry of Education's five types of higher education institutions, which are listed in descending order of their on-time program duration.[48] Column (B) shows the number of exit exam takers per year. The majority of exam takers are from university-level institutions, with fewer students from technical colleges.

Column (C) shows the fraction of these 310 colleges that appear in the Ministry of Education records that I use in my analysis. These proportions are weighted by the number

[48] Most programs at universities require 4–5 years of study, while programs at Technical/Professional Institutes typically take 2–3 years.

Table C3. Higher education institutions in Ministry of Education records

	(A) Number of colleges	(B) Number of exit exam takers/year	(C) Prop. of colleges in records
University	122	134,496	1.00
University Institute	103	53,338	0.88
Technology School	3	2,041	1.00
Technology Institute	47	15,092	0.82
Technical/Professional Institute	35	11,408	0.99
Total	310	216,375	0.96

Notes: Column (A) depicts the number of colleges that have *Saber Pro* exit exam takers in 2009–2011 using administrative records from the testing agency. Colleges are categorized into the Ministry of Education's five higher education institution types. Column (B) shows the number of 2009–2011 exam takers per year. Column (C) shows the proportion of colleges that appear in the Ministry of Education records, where colleges are weighted by the number of exit exam takers.

of exam takers depicted in column (B). Column (C) shows that the Ministry of Education records include all universities but are missing a few technical colleges.[49] Overall, 96 percent of exit exam takers attend colleges that appear in the Ministry of Education records.

I define my measure of college quality, Q_c, as a college's percentile rank in each exam cohort based on the mean pre-reform exam score in each college c.[50] For the 20 colleges with fewer than ten pre-reform enrollees, I define Q_c as the mean exam percentile of the students at all 20 of these colleges.

My last data source is from the Ministry of Social Protection. These data provide monthly earnings for any college enrollee employed in the formal sector in 2008–2012. From these records I calculate the log of an individual's average daily earnings measured 10–11 years after taking the admission exam. Ten and eleven years after the test are the two experience levels at which I can measure earnings for each of the 1998–2001 exam cohorts. I compute average daily earnings by dividing total annual earnings by the number of formal employment days in the year, demeaning by exam cohort and year, and averaging across the two years.

I merge these three datasets using national ID numbers, birth dates, and names. Nearly all students in these records have national ID numbers, but Colombians change ID numbers around age 17. Most students in the admission exam records have the below-17 ID number (*tarjeta*), while the majority of students in the college enrollment and earnings records have the above-17 ID number (*cédula*). Merging using ID numbers alone would therefore lose a large majority of students. Instead, I merge observations with either: 1) the same ID number

[49] The largest omitted institutions are the national police academy (*Dirección Nacional de Escuelas*) and the Ministry of Labor's national training service (*Servicio Nacional de Aprendizaje*).

[50] This measure is the average score across all exam subjects.

and a fuzzy name match; 2) the same birth date and a fuzzy name match; or 3) an exact name match for a name that is unique in both records.

38 percent of the 1998–2001 exam takers appear in the enrollment records, which is comparable to the higher education enrollment rate in Colombia during the same time period.[51] A better indicator of merge success is the percentage of college enrollees that appear in the admission exam records because all domestic college students must take the exam. I match 88 percent of enrollees who took the admission exam between 1998 and 2001.[52]

The exam validity results in Table 3 use college GPA computed from transcript records from one public flagship university (name withheld). I obtained transcript records for students in business, engineering, and architecture programs in the 2000–2004 enrollment cohorts. I calculate first-year GPA as the average grade across all courses the student took in the first two semesters after enrolling.

C.2. **The 2000 admission exam reform.** This section provides further details on the 2000 reform of the national Colombian college admission exam.

The goal of the 2000 exam overhaul was to design an exam that supported the dual goals of measuring high school quality and aiding in college admissions. The pre-reform exam was thought to primarily test intellectual ability and rote memorization, and was thus poorly suited for measuring the contribution of high schools to students' educational development. Furthermore, the exam was criticized for being biased toward certain students depending on their gender or family background.

To achieve this goal, the testing agency rewrote the exam with the aim of testing "competencies" rather than "content." The focus of the new was to test "know-how in context," which means that students should be able to apply a given piece of information to different situations. Examples of such competencies include interpreting a text, graphic, or map in solving a problem, and assessing different concepts and theories that support a decision. The post-reform exam therefore placed a greater emphasis on communication skills, as it asked students to interpret, argue, and defend their answers.

[51] The gross tertiary enrollment rate ranged from 22 percent to 24 percent between 1998 and 2001 (World Bank World Development Indicators, available at http://data.worldbank.org/country/colombia in October 2016). This rate is not directly comparable to my merge rate because not all high school aged Colombians take the ICFES exam. About 70 percent of the secondary school aged population was enrolled in high school in this period. Dividing the tertiary enrollment ratio by the secondary enrollment ratio gives a number roughly comparable to my 38 percent merge rate.

[52] The enrollment records contain age at time of the admission exam for some students, which allows me to calculate the year they took the exam. Approximately 16 percent of students in the enrollment dataset have missing birth dates, which accounts for the majority of observations I cannot merge. Some duplicate matches arise because students took the admission exam more than once, though I erroneously match a small number of students with the same birth date and similar names.

Table C4. Mean admission score by exam component and cohort

Subject groups	Exam components	1998	1999	2000	2001
		Exam cohort			
Biology	Biology	48.0	48.3	45.1	44.6
Chemistry	Chemistry	45.7	50.2	44.8	45.0
Language	Language	48.8	50.8	46.4	46.4
Math	Math aptitude	49.1	50.4		
	Math knowledge	48.8	48.5		
	Math			43.1	41.2
Physics	Physics	47.1	46.7	45.2	46.7
S. sciences	Social sciences	48.1	48.6		
	Geography			44.6	43.5
	History			43.7	43.5
Excluded components	Verbal aptitude	48.6			
	Philosophy			44.7	43.7
	Foreign language			40.9	42.0
	Elective	49.7	50.9	51.9	55.3
	Mean (all components)	48.2	49.3	45.0	45.2
	St. dev. (all components)	10.2	10.2	7.4	7.5

Notes: The sample includes all exam takers in the ICFES records.

Figures C1–C4 present sample questions from the biology, language, math, and social sciences components of the pre-reform and post-reform exams. These sample questions were distributed by the testing agency to describe the central motivation of the overhaul. Questions from the pre-reform exam are briefer and require more memorization. The post-reform sample questions are longer and often include a figure or passage that the student must interpret. Further, some pre-reform questions have a complicated answer structure, while the post-reform questions are all straightforward multiple choice.

These communication and interpretation skills were tested in the context of subjects from the core secondary education curriculum. To better align the test with the high school curriculum, the reform also altered the specific subjects that were tested. Table C4 shows the subject components that were included in the admission exam between 1998 and 2001. The 2000 reform combined two math exams—one designed to measure aptitude and another designed to test knowledge—into a single component. The reform also split the social sciences component into separate tests for history and geography. Further, the 2000 reform added components in philosophy and foreign language, which was English for the large majority of students.

In Section 2 I focus on the six subject groups listed in the leftmost column of Table C4: biology, chemistry, language, math, physics, and social sciences. I average the pre-reform math aptitude and math knowledge components into a single math score. I also average the

Panel A. Pre-reform sample question

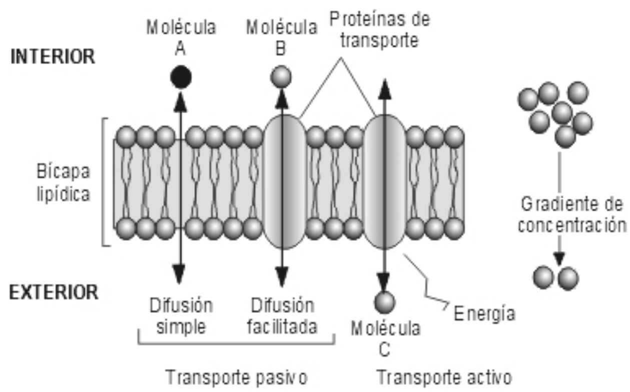
Which of the following two are inverse chemical processes?

1. Photosynthesis
2. Cyclosis
3. Breathing
4. Circulation

(A) If 1 and 2 are correct, fill in oval A
(B) If 2 and 3 are correct, fill in oval B
(C) If 3 and 4 are correct, fill in oval C
(D) If 2 and 4 are correct, fill in oval D
(E) *If 1 and 3 are correct, fill in oval E*

Panel B. Post-reform sample question

The diagram shows a cell that is exchanging substances with its environment through the cell membrane.



If at a certain time it is observed that the number of molecules A entering the cell is greater than the number coming out of it, it can be assumed that within the cell there is

(A) A higher concentration of molecules than outside
(B) *A lower concentration of molecules than outside*
(C) A molecule concentration equal to that outside
(D) An absence of molecules A

Figure C1. Biology sample questions

Notes: Correct answers are in *italics*. The sample question from the pre-reform exam was obtained from a version of the ICFES testing agency's website that was archived in January 1997 (available at https://web.archive.org/web/19980418191357/http://acuario.icfes.gov.co/12/122/1222/12223/Tipos.html in October 2016). The sample question from the post-reform exam was obtained from a September 2008 ICFES report entitled "State Assessment Tests in Colombia" (*"Evaluación con Pruebas de Estado en Colombia"*) (available at http://www.ieia.com.mx/materialesreuniones/1aReunionInternacionaldeEvaluacion/PONENCIAS18Septiembre/ConferenciasMagnas/MargaritaPenaBorrero.pdf in October 2016).

post-reform history and geography components into a single social sciences score. I exclude the verbal component, which appears only in the pre-reform exam, and the philosophy and foreign language components, which appear only in the post-reform exam. I also exclude the elective component, which was rarely used by colleges to determine admissions.

Panel A. Pre-reform sample question

The phrase:

"¿Estará Pedro en la casa?"

is used to ask about the location of Pedro:

(A) *At the moment when the question is asked*
(B) At a future moment
(C) At any moment
(D) At the moment when the answer is given

Panel B. Post-reform sample question

Me parece que no es preciso demostrar que la novela policial es popular, porque esa popularidad es tan flagrante que no requiere demostración. Para explicarla—aquellos que niegan al género su significación artística—se fundan en la evidencia de que la novela policial ha sido y es uno de los productos predilectos de la llamada "cultura de masas," propia de la moderna sociedad capitalista.

La popularidad de la novela policial sería, entonces, sólo un resultado de la manipulación del gusto, sólo el fruto de su homogeneización mediante la reiteración de esquemas seudoartísticos, fácilmente asimilables, y desprovistos, claro, de verdadera significación gnoseológica y estética; sazonados, además, con un puñado de ingredientes de mala ley: violencia, morbo, pornografía, etcétera, productos que se cargan, casi siempre, de mistificaciones y perversiones ideológicas, tendientes a la afirmación del estatus burgués y a combatir las ideas revolucionarias y progresistas del modo más burdo e impúdico.

Pero hay que decir que ello constituye no sólo una manipulación del gusto en general, sino también una manipulación de la propia novela policial, de sus válidas y legítimas manifestaciones, una prostitución de sus mecanismos expresivos y sus temas. Los auténticos conformadores del género policial (no hay que olvidarlo) fueron artistas de la talla de Edgar Allan Poe y Wilkie Collins. Y desde sus orígenes hasta nuestros días, el género ha producido una buena porción de obras maestras.

From "La novela policial y la polémica del elitismo y comercialismo"
In *Ensayos Voluntarios*, Guillermo Rodríguez Rivera.
Havana, *Editorial Letras Cubanas*, 1984.

The theme of the previous text is:

(A) The pseudo-artistic nature of detective novels is devoid of epistemological and aesthetic significance
(B) The detective novel is a favorite product of the so-called "mass culture"
(C) The popularity of the detective genre is not necessary to show through evidence
(D) *Detective novels and their manifestations can manipulate tastes*

FIGURE C2. Language sample questions

Notes: Correct answers are in *italics*. The sample question from the pre-reform exam was obtained from a version of the ICFES testing agency's website that was archived in January 1997 (available at https://web.archive.org/web/19980418191357/http://acuario.icfes.gov.co/12/122/1222/12223/Tipos.html in October 2016). The sample question from the post-reform exam was obtained from a September 2008 ICFES report entitled "State Assessment Tests in Colombia" (*"Evaluación con Pruebas de Estado en Colombia"*) (available at http://www.ieia.com.mx/materialesreuniones/1aReunionInternacionaldeEvaluacion/PONENCIAS18Septiembre/ConferenciasMagnas/MargaritaPenaBorrero.pdf in October 2016).

Panel A. Pre-reform sample question

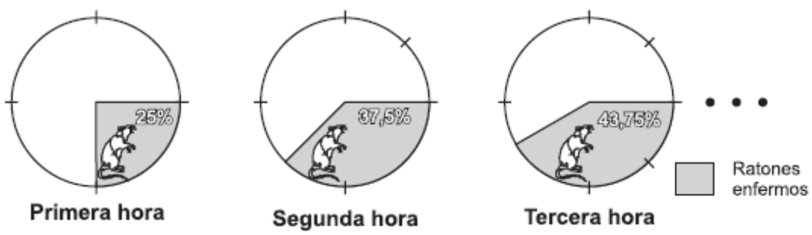
It is known that the result of multiplying a number by itself several times is 256. You can identify this number if it is known

I. Whether the number is positive or negative
II. How many times the number is multiplied by itself

(A) If fact I is enough to solve the problem, but fact II is not, fill in oval A
(B) If fact II is enough to solve the problem, but fact I is not, fill in oval B
(C) *If facts I and II together are sufficient to solve the problem, but each separately it is not, fill in oval C*
(D) If each of facts I and II separately are sufficient to solve the problem, fill in oval D
(E) If facts I and II together are not enough to solve the problem, fill in oval E

Panel B. Post-reform sample question

To test the effect of a vaccine applied to 516 healthy mice, an experiment was performed in a laboratory. The goal of the experiment is to identify the percentage of mice that become sick when subsequently exposed to a virus that attacks the vaccine. The following graphs represent the percentage of sick mice after the first, second, and third hours of the experiment.



With regard to the state of the mice, it is NOT correct to say that

(A) *After the first hour there are only 75 healthy mice*
(B) After the first hour there are 129 sick mice
(C) After two and a half hours there are more healthy mice than sick mice
(D) Between the second and third hour the number of sick mice increased by 6.25 percentage points

FIGURE C3. Math sample questions

Notes: Correct answers are in *italics*. The sample question from the pre-reform exam was obtained from a version of the ICFES testing agency's website that was archived in January 1997 (available at https://web.archive.org/web/19980418191357/http://acuario.icfes.gov.co/12/122/1222/12223/Tipos.html in October 2016). The sample question from the post-reform exam was obtained from a September 2008 ICFES report entitled "State Assessment Tests in Colombia" (*"Evaluación con Pruebas de Estado en Colombia"*) (available at http://www.ieia.com.mx/materialesreuniones/1aReunionInternacionaldeEvaluacion/PONENCIAS18Septiembre/ConferenciasMagnas/MargaritaPenaBorrero.pdf in October 2016).

Table C4 also shows that the reform affected the mean and the variance of exam scores. The bottom rows show that the mean score across all subjects was near 50 in the pre-reform cohorts, and about 45 in the post-reform cohorts. Further, the standard deviation across all components fell from approximately ten to 7.5. My interest is in students' relative performance, and so for all my analysis I normalize exam scores to percentiles within each exam cohort.

Panel A. Pre-reform sample question

<u>Assertion</u>: The only factor that determined the abolition of slavery in Colombia in the mid-nineteenth century was the economy.

<u>Reason</u>: In the mid-nineteenth century the formation of regional markets and the development of agriculture in our country made it necessary to establish freedom of labor.

(A) If the assertion and reason are true and the reason is a correct explanation of the claim, fill in oval A
(B) If the assertion and reason are true, but the reason is not a correct explanation of the claim, fill in oval B
(C) If the assertion is true but the reason is a false proposition, fill in oval C
(D) *If the assertion is false but the reason is a true proposition, fill in oval D*
(E) If both assertion and reason are false propositions, fill in oval E

Panel B. Post-reform sample question

In South America, archaeological finds of pottery—used for food preparation and storage of grain—have been interpreted as evidence of the strengthening of agriculture between the Andean cultures before the Inca Empire. These findings are indicative of agricultural and sedentary cultures because

(A) They reflect the broad expanse of corn, cacao, and vegetables
(B) There are no findings of hunting weapons made of stone
(C) *Nomadic activities, in contrast, require little ceramic production*
(D) Large irrigation systems are part of the same findings

FIGURE C4. Social sciences sample questions

Notes: Correct answers are in *italics*. The sample question from the pre-reform exam was obtained from a version of the ICFES testing agency's website that was archived in January 1997 (available at https://web.archive.org/web/19980418191357/http://acuario.icfes.gov.co/12/122/1222/12223/Tipos.html in October 2016). The sample question from the post-reform exam was obtained from a September 2008 ICFES report entitled "State Assessment Tests in Colombia" (*"Evaluación con Pruebas de Estado en Colombia"*) (available at http://www.ieia.com.mx/materialesreuniones/1aReunionInternacionaldeEvaluacion/PONENCIAS18Septiembre/ConferenciasMagnas/MargaritaPenaBorrero.pdf in October 2016).

C.3. **Derivation of differences-in-differences specifications.** This section derives my benchmark differences-in-differences specifications (equations (7) and (8)) from the key model equations in Section 1.

First consider an equation that defines the socioeconomic test score gap,

$$T_{it} = \psi_t + \theta_t^T X_i + u_{it}, \qquad \text{(C1)}$$

where T_{it} is the test score percentile for student i in exam cohort t, and X_i is an indicator for high SES individuals as defined in Section 2.2. Equation (C1) is analogous to the definition of the test score gap in equation (1) from Section 1, but it estimates separate gaps for each exam cohort. The intercept, ψ_t, is the average test score for low SES students in cohort t, and $\theta_t^T = \bar{T}_{1t} - \bar{T}_{0t}$ gives the SES exam score gap in cohort t. The results in Section 2.4 show that test score gaps, θ_t^T, declined with the exam reform.

To incorporate geographic variation in exam stakes, modify the match function (4) to allow for different admission effects of test scores in each geographic area g:

$$Q_c = \phi_g T_i + u_{icg},\psi \tag{C2}$$

where Q_c measures college quality. As described in Section 2.2, I define Q_c to be a college's percentile rank based on the mean pre-reform admission exam score of its student body. This follows the common practice of using mean test scores to measure school quality (Dale and Krueger, 2002; Hoxby, 2009). The coefficient $\phi_g\psi$ thus gives the effect of individual test scores on college mean test scores in area g. The discussion in Section 3.2 implies that the $\phi_g\psi$ coefficients are larger in treated areas.

Plugging equation (C1) into equation (C2) and renaming coefficients yields

$$Q_c = {}_{gt} + \theta_{gt} X_i + u_{icgt}.\psi \tag{C3}$$

The intercept in this regression, ${}_{gt}$, gives the average college quality for low SES students ($X_i = 0$) in area $g\psi$ and exam cohort t. The slope coefficient, θ_{gt}, gives the SES "college quality gap" in each geographic area/cohort pair. The main prediction is that θ_{gt} should decline in treated areas relative to control areas with the exam reform. I capture this effect in a single coefficient using a standard differences-in-differences regression, where the dependent variable is the college quality gap, θ_{gt}:

$$\theta_{gt} = {}_g + {}_t + \theta(\text{Treated}_g \times \text{Post}_t) + u_{gt}.\psi \tag{C4}$$

This regression includes geographic area dummies, ${}_g$, and exam cohort dummies, ${}_t$. The variable of interest is the interaction the treatment variable, Treated_g, with a dummy for post-reform cohorts, Post_t. The coefficient of interest, θ, measures the average change in the college quality gap in treated areas relative to control areas.

Plugging equation (C4) into equation (C3) yields my first stage empirical specification (8):

$$Q_c = {}_{gt} + \Big({}_g + {}_t + \theta^q (\text{Treated}_g \times \text{Post}_t)\Big) X_i + u_{icgt}.\psi \tag{C5}$$

Equation (C5) differs from a standard differences-in-differences regression in that it measures changes in a slope rather than changes in levels. In this case, the "slope" is the SES gap in college quality. The main prediction is $\theta^q < \psi 0$, i.e., the college quality gap declines more in treated areas than in control areas with the exam reform. In other words, in treated areas the reform should shift low SES students into higher quality colleges and displace high SES students to lower quality colleges.

To measure effects of the admission exam reform on outcomes, I use a standard differences-in-differences regression in levels (equation (7)):

$$\text{(C6)} \qquad y_{igt} = \ _g + \ _t + \theta(\text{Treated}_g \times \text{Post}_t) + u_{igt}.\psi$$

This regression is analogous to specification (C4) but it uses different dependent variables, y_{igt}. The variable of interest is the interaction between indicators for treated areas and post-reform cohorts. The main outcome is log earnings measured 10–11 years after the admission exam, and in this case the coefficient $\theta\psi$ measures the reform's effect on average earnings in treated areas relative to control areas. In addition, I use indicators for college persistence and graduation as dependent variables to explore the mechanisms underlying earnings effects.

C.4. **Magnitude of college quality effects.** This section describes the magnitude of the reform's effect on the relationship between SES and college quality.

The estimates in the first row of Table 5 suggest a roughly two percentile point decrease in the college quality gap between top and bottom income quartile students. Treated regions have about five colleges on average. Thus the estimate is equivalent to one in 20 low-income students moving up a college quality tier (a one percentile point increase in mean college quality), and one in 20 high-income students moving down a college tier (a one percentile point decrease in mean college quality). In total, there were roughly 20,000 bottom income quartile college enrollees per cohort in treated regions.[53] This implies that roughly 1,000 low-income students moved up a college tier.

These 1,000 low-income students were spread out over more than 100 treated region colleges and multiple enrollment cohorts. Thus it is unlikely that colleges experienced a perceptible change in the SES composition of their students in the initial years after the reform. This argues against peer effects or institutional responses as a potential mechanism for the graduation and earnings effects (Table 6).

C.5. **Estimation of causal and instrumental forests.** This section gives details on the estimation of first stage effects and returns to college quality using Athey et al. (2017)'s causal and instrumental forest methods.

I begin by computing the variables $\tilde{w}_i$, $\tilde{Q}_c$, and $\widetilde{(\text{Treated}_g \times \text{Post}_t)}$, which are residuals from regressions of log earnings, college quality, and treatment status on region dummies and cohort dummies.[54] I divide college quality by ten so that one unit equals ten percentile points. With these residuals, the goal is to estimate heterogeneity in the first stage effects

[53] From Table 4, there are 20 treated regions with 11,500 exam takers per cohort on average. The mean college enrollment rate in treated regions is 34 percent. Thus there were roughly 80,000 college enrollees from treated region in each cohort, including 20,000 from the bottom income quartile.

[54] For the results in Appendix Table A17, I use the municipality-level treatment variable and municipality dummies.

θ_z and returns to college quality $\ _z$ for covariate values $Z_i = z\psi$using the 2SLS moment equations

$$E\psi\left[\left(\widetilde{\text{Treated}_g \times \text{Post}_t}\right)\left(\tilde{Q}_c - \theta_z(\widetilde{\text{Treated}_g \times \text{Post}_t})\right)\right] = 0,\psi \tag{C7}$$

$$E\psi\left[\left(\widetilde{\text{Treated}_g \times \text{Post}_t}\right)\left(\tilde{w}_i - \ _z\tilde{Q}_c\right)\right] = 0.\psi \tag{C8}$$

Table C5 shows the 28 variables I include in the covariate vector Z_i. I define all covariates as binary variables and ensure that the mean of each variable is 0.2 or above.[55] I drop students without formal sector earnings, which reduces the sample size from 612,949 (column (B), Table 1) to 340,623. I also drop students with missing values of any covariate, which reduces the sample to 270,400 observations.

I divide the sample randomly in half into a training sample and a validation sample. I estimate moment (C7) on the training sample using the `causal_forest` function from Athey et al. (2017)'s `grf` package for R. I estimate moments (C7)–(C8) on the training sample using the `instrumental_forest` function. I use the same seed and grow 2,000 trees for each estimation.

The main implementation choices for the `grf` functions are the minimum number of observations in each node (`nodesize`) and the maximum imbalance of a split (`alpha`). These options affect the minimum allowable size of covariate groups $z\psi$and the variable values that can be combined to form these groups. Using covariate groups that are too large limits the ability to estimate heterogeneity in $\ _z$, while groups that are too small reduce the precision of these estimates.

I choose the values of `nodesize` and `alpha` that maximize the sum of the $F\psi$statistics from tests that each $\ _z$ is equal to the mean of the $\ _z$ estimates. This balances the above considerations by rewarding both the number of $\ _z$ estimates as well as their precision. Table C6 shows results from the tuning of these parameters and the selected values (`nodesize` = 1,000 and `alpha` = 0.15).

Lastly, I predict the parameters θ_z and $\ _z$ estimated using these options into the validation dataset. I use this dataset for Table 7 (Panel B) and the results in Section 5.5. Columns (A) and (B) in Table C7 show the variable importance percentages for each covariate, Z_i, from the causal and instrumental forest estimations. Column (C) shows the correlation between the estimated returns to college quality, $\ _z$, and each covariate using the sample from Column (D) in Table 7 (Panel B). Column (D) shows these same correlations in the subsample of students from the bottom income quartile.

[55] In estimation, continuous covariates—or binary covariates with infrequently observed characteristics—lead to imprecisely estimated treatment effects.

TABLE C5. Definition of variables in covariate vector Z_i

#	Variable	Definition	Mean
1.	Female	=1 if female	0.53
2.	Mother's education (low)	=1 if primary education or less	0.38
3.	Mother's education (middle)	=1 if secondary education	0.36
4.	Mother's education (high)	=1 if some college or more	0.27
5.	High school rank (low)	=1 if inferior or low rank	0.29
6.	High school rank (middle)	=1 if middle rank	0.35
7.	High school rank (high)	=1 if high, superior, or far superior rank	0.36
8.	High school academic level	=1 if academic high school	0.51
9.	High school ownership	=1 if high school is public	0.55
10.	Age	=1 if 17 or younger	0.56
11.	Father's education (low)	=1 if primary education or less	0.38
12.	Father's education (middle)	=1 if secondary education	0.28
13.	Father's education (high)	=1 if some college or more	0.34
14.	Number of siblings (low)	=1 if zero or one sibling	0.28
15.	Number of siblings (middle)	=1 if two siblings	0.32
16.	Number of siblings (high)	=1 if three or more siblings	0.40
17.	Father's occupation	=1 if father is not a laborer/unemployed*	0.41
18.	Birth order (low)	=1 if first born	0.36
19.	Birth order (middle)	=1 if second born	0.29
20.	Birth order (high)	=1 if third born or more	0.34
21.	Number of siblings with/in higher ed. (low)	=1 if zero	0.50
22.	Number of siblings with/in higher ed. (middle)	=1 if one	0.30
23.	Number of siblings with/in higher ed. (high)	=1 if two or more	0.20
24.	Number of family members employed	=1 if two family members employed	0.41
25.	Family income (bottom quartile)	=1 if family income in quartile 1	0.24
26.	Family income (quartile 2)	=1 if family income in quartile 2	0.25
27.	Family income (quartile 3)	=1 if family income in quartile 3	0.25
28.	Family income (top quartile)	=1 if family income in quartile 4	0.26

Notes: Means are computed from the validation sample ($N = 135,200$).

* Father's occupations classified as not laborer/unemployed include entrepreneur, director/manager, independent professional, employed professional, employed worker, retired, or student. Laborer/unemployed occupations include independent worker, annuitant, labor, home, no income/looking for work, and deceased.

Table C6. Tuning of `nodesize` and `alpha` parameters

(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)
		All students			Bottom income quartile			
`nodesize`	`alpha`	N with $F_\theta > 10$	# z cells	Sum of F: ${}_z = \bar{\ }_z$	N with $F_\theta > 10$	# z cells	Sum of F: ${}_z = \bar{\ }_z$	Selected
250	0.05	2,935	522	653.5	169	48	26.9	
250	0.10	3,436	639	674.8	263	58	35.9	
250	0.15	2,845	525	500.5	321	63	57.5	
250	0.20	3,204	574	864.3	332	54	43.7	
500	0.05	4,531	741	1,102.4	574	82	44.0	
500	0.10	4,251	765	1,092.4	301	69	47.2	
500	0.15	3,359	679	798.1	348	78	65.2	
500	0.20	3,684	691	1,226.4	314	77	48.3	
1,000	0.05	4,821	927	1,345.7	573	122	69.1	
1,000	0.10	4,801	924	1,638.4	510	121	54.1	
1,000	0.15	4,548	947	1,699.0	516	127	154.8	✓
1,000	0.20	4,111	793	1,541.6	375	145	101.6	
2,500	0.05	2,664	548	945.7	439	120	55.5	
2,500	0.10	2,722	310	1,183.2	175	44	8.8	
2,500	0.15	2,904	574	1,259.6	252	66	20.9	
2,500	0.20	4,633	789	1,335.8	487	96	54.7	
5,000	0.05	2,513	793	1,072.5	157	60	14.3	
5,000	0.10	3,715	1,007	371.7	636	141	35.2	
5,000	0.15	3,343	799	714.0	385	89	39.0	
5,000	0.20	3,281	819	491.2	241	82	25.9	

Notes: Columns (A) and (B) show the `nodesize` and `alpha` parameters used to estimate the `causal_forest` and `instrumental_forest` functions in Athey et al. (2017)'s `grf` package for R.

Column (C) shows the number of students in the validation sample who have first stage F statistics greater than ten. Column (D) shows the number of unique covariate cells z in this subsample. Column (E) shows the sum of the F statistics from tests that each return to college quality, ${}_z$, is equal to the mean return.

Columns (F)–(H) are analogous to columns (C)–(E), but these statistics are from the subsample of students with bottom quartile family incomes. Column (I) shows the tuning parameters used for the results in Section 5.5, which are selected to maximize column (E).

Table C7. Covariate variable importance and correlations with ${}_z$

		(A)	(B)	(C)	(D)	(E)	(F)
		Variable importance		Correlations with θ_z		Correlations with ${}_z$	
#	Variable	θ_z	${}_z$	All students	Bottom income Q	All students	Bottom income Q
1.	Female	0.016	0.035	0.03	0.14	0.04	0.29
2.	Mother's education (low)	0.014	0.015	0.22	-0.15	-0.34	-0.18
3.	Mother's education (middle)	0.039	0.033	0.24	0.17	-0.49	0.16
4.	Mother's education (high)	0.016	0.025	-0.51	-0.04	0.74	0.08
5.	High school rank (low)	0.023	0.037	0.34	0.40	-0.59	-0.64
6.	High school rank (middle)	0.024	0.041	0.22	-0.23	-0.44	0.52
7.	High school rank (high)	0.109	0.091	-0.54	-0.33	0.89	0.42
8.	High school academic level	0.033	0.038	-0.24	-0.13	0.58	0.24
9.	High school ownership	0.014	0.022	0.34	0.00	-0.53	0.30
10.	Age	0.059	0.047	-0.26	-0.25	0.52	-0.12
11.	Father's education (low)	0.007	0.010	0.20	-0.16	-0.21	-0.11
12.	Father's education (middle)	0.054	0.045	0.26	0.29	-0.62	0.06
13.	Father's education (high)	0.041	0.038	-0.45	-0.15	0.77	0.27
14.	Number of siblings (low)	0.009	0.016	-0.16	-0.06	0.42	0.09
15.	Number of siblings (middle)	0.024	0.023	-0.13	-0.31	0.02	0.03
16.	Number of siblings (high)	0.025	0.028	0.27	0.33	-0.44	-0.10
17.	Father's occupation	0.035	0.027	-0.09	-0.12	-0.25	-0.29
18.	Birth order (low)	0.012	0.025	-0.11	-0.14	0.11	-0.18
19.	Birth order (middle)	0.011	0.018	-0.09	-0.24	0.14	0.05
20.	Birth order (high)	0.018	0.024	0.19	0.34	-0.27	0.13
21.	No. siblings with/in higher ed. (low)	0.012	0.013	-0.04	-0.29	0.05	0.04
22.	No. siblings with/in higher ed. (middle)	0.014	0.014	-0.02	0.16	0.02	0.03
23.	No. siblings with/in higher ed. (high)	0.008	0.013	0.07	0.19	-0.10	-0.09
24.	Number of family members employed	0.017	0.021	-0.04	0.08	-0.08	-0.01
25.	Family income (bottom quartile)	0.014	0.020	0.21		-0.37	
26.	Family income (quartile 2)	0.114	0.088	0.43		-0.62	
27.	Family income (quartile 3)	0.011	0.016	0.06		-0.15	
28.	Family income (top quartile)	0.226	0.177	-0.68		0.88	

Notes: Column (A) shows the variable importance percentage for each covariate Z_i from the causal forest estimation of the first stage effects, θ_z. Column (B) shows the variable importance percentages from the instrumental forest estimation of the returns to college quality, ${}_z$.

Column (C) shows the correlations between each covariate Z_i and θ_z for the sample from column (A) in Table 7 (Panel B). Column (D) shows the same correlations for the subset of students with bottom quartile incomes.

Column (E) shows the correlations between each covariate Z_i and ${}_z$ for the sample from column (D) in Table 7 (Panel B). Column (F) shows the same correlations for the subset of students with bottom quartile incomes.