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

EVAN RIEHL*

MARCH 2018

ABSTRACT. This paper asks whether socioeconomic gaps in college admission test scores also translate into greater earnings inequality. I show that the link between test scores and labor market outcomes depends on how well the exam measures students' potential earnings returns to college quality. I estimate the distribution of these returns by exploiting a natural experiment from a redesign of the Colombian national college admission exam. I find substantial heterogeneity in the returns to college quality including negative returns for some low-income students. This suggests that raising low-income students' test scores may reduce their future earnings if the exam cannot identify which students would benefit from attending a better college.

* 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 Bergman, 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.

1. Introduction

Are standardized tests a fair way to allocate college admissions? In nearly all standardized exams a student's performance is strongly related to her parents' income, resulting in large test score gaps between socioeconomic groups. This leads many to argue that admission tests should be redesigned, or eliminated altogether, to increase socioeconomic diversity at top colleges.

This paper explores the question of fairness from a different perspective—how should a college admission test be designed if one cares about inequality in labor market outcomes rather than inequality in admissions? I develop a framework to show that the relationship between test score gaps and earnings inequality depends on how the exam relates to students' *returns to college quality*—their earnings benefits to attending a better school. If these returns are positive and vary little across students, reducing the socioeconomic gap in test scores lowers earnings inequality regardless of the design of the exam. If instead returns to college quality are heterogeneous, reducing the test score gap can harm disadvantaged students on the labor market if the exam becomes a worse measure of this heterogeneity.

I then estimate the distribution of these returns using administrative data from the country of Colombia that links students' admission test scores to their college and labor market outcomes. For identification, I exploit a reform of the national college admission exam that dramatically reduced the socioeconomic test score gap and shifted low-income students into higher quality colleges. I find substantial heterogeneity in the return to college quality including negative returns for some low-income students. The reform reduced the exam's predictive power for these returns, and as a result lowered average earnings both for disadvantaged students and for the student population as a whole. Thus the fairness of a college admission exam may depend not just on how its scores correlate with socioeconomic status, but also on whether it identifies students who would benefit from attending a better college.

The first part of the paper develops a framework to illustrate how the socioeconomic test score gap affects the distribution of post-college earnings. The model considers a population of students defined by two characteristics: their socioeconomic status (SES), and a vector of abilities that reflect their capacity to correctly answer potential questions on an exam. The role of the testing agency is to choose a vector of ability weights that yield a single normalized test score for each student. By assigning higher weights to abilities that are less correlated with SES, the testing agency can reduce the socioeconomic gap in test scores. There may be a tradeoff in reducing test score gaps, however, if the exam becomes a worse measure of abilities that predict college success.

The key parameter in the model is the return to college quality, which measures a student's potential benefits from attending a better college. I assume students are sorted into colleges

according to their performance on the admission test, and define a college's quality as the mean test score of its student body (Dale and Krueger, 2002; Hoxby, 2009). The return to college quality, which I denote by β_i, is the coefficient that relates student i's potential post-college earnings to the quality of the college they attend. The i subscript on this coefficient allows for heterogeneity in these returns across students.

The labor market implications of the SES test score gap depend on how much heterogeneity there is in the return to college quality, and on how well the exam measures this heterogeneity. If the return to college quality is positive and constant for all individuals ($\beta_i = \beta > 0$), then reducing the test score gap can lower the SES earnings gap without changing average earnings in the market. Intuitively, earnings gains to low SES students are exactly offset by losses to high SES students when β_i is constant. If instead β_i is correlated with SES, then there can be an equity/efficiency tradeoff in reducing the test score gap.

Further, under some conditions on the returns to college quality, lowering the test score gap can actually harm low SES students. If the average value of β_i is negative for low SES students, reducing the test score gap can lower their average earnings by shifting some students into colleges where they are unlikely to succeed. This is a version of the "mismatch hypothesis," which states that some disadvantaged students may be better off at lower ranked colleges (Arcidiacono and Lovenheim, 2016). Reducing the test score gap can also decrease low SES earnings if there is substantial heterogeneity in β_i and the new test becomes a worse measure of this variation. In this case, there is "mismatch" because students with the highest values of β_i are less likely to gain admission to the best colleges.

The framework shows that the implications of test score gaps for earnings inequality depend on heterogeneity in the returns to college quality. In the rest of the paper, I exploit data and a natural experiment from the country of Colombia to estimate the distribution of 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 exam.

To analyze the effects of the Colombian reform, I link individual-level administrative records that contain admission exam performance, college outcomes, and early career earnings for the near universe of college students in the country. With these data I measure the reform's effects on three different outcomes: 1) the SES gap in test scores; 2) the SES gap in college quality; and 3) earnings after college. Lastly, I combine the college quality and earnings effects in an instrumental variables approach to identify distributional characteristics of the return to college quality, β_i.

First, I show that the Colombian reform led to a dramatic reduction in the socioeconomic test score gap, but at the cost of becoming a worse predictor of students' college outcomes. Across various measures of socioeconomic status, the gap in average test scores between

high SES and low SES students fell sharply in the first cohort that took the new exam. For example, in several subjects the gap between students in the top and bottom quartiles of the parental income distribution decreased by about 50 percent. However, the validity of the admission exam—as measured by its correlation with college grades, test scores, and graduation rates—also dropped sharply with the reform.

Second, I show that the reform reduced the SES "college quality gap" in regions where the national test was important for admission to top colleges. College attendance in Colombia is highly localized, and each region of the country has a large flagship public university that is often its most selective college. In regions that I call the "treated" areas, the flagship university admits students solely based on scores from the national test. In "control" regions, flagships admissions are based on the university's own exam. I use this regional variation in a differences in differences design to show that the reform caused a reduction in the SES college quality gap as measured by school mean pre-reform admission scores. That is, low SES students in treated regions became more likely to attend high quality colleges relative to their peers in controls regions. Correspondingly, high SES students in treated regions were shifted into lower quality colleges.

Third, I show that the reduction the college quality gap lead to a market-wide decline in earnings, with negative effects for both high SES and low SES students. Specifically, average earnings measured one decade after the admission exam fell by about 1.5 percent in treated regions relative to control regions. This decline in earnings came from high SES students, who were displaced to lower quality colleges, but also from low SES students, who were shifted into higher quality colleges. These results hold in differences in differences regressions and in event-study graphs, which show sharp decreases in outcomes in the first cohort after the reform. I also show that college graduation rates declined for both high and low SES students, suggesting that the exam reform caused students to drop out of college and enter the labor market with lower educational credentials.

Lastly, I use these results to develop an instrumental variables approach that estimates heterogeneity in the returns to college quality. I create instrumental variables based on the differences in differences specification that identify the average return to college quality for different SES groups. I also show how new techniques that use random forests to estimate heterogeneity in causal effects (Athey et al., 2016) can identify variation in these returns within SES groups.

The main result from the instrumental variables analysis is that there is substantial heterogeneity in returns to college quality, and the Colombian reform reduced the exam's predictive power for this heterogeneity. For the students most affected by the reform, returns to college quality are strongly correlated with SES, including negative returns for some low SES students. On average, a ten percentile point in college quality led to a five percent increase in

earnings for high SES students, and a five percent *decrease* in earnings for low SES students. The returns to college quality also vary substantially within SES groups, with an overall standard deviation of about seven log points. I find that the reform reduced the correlation between admission exam scores and these returns, including a decline in predictive power both across and within SES groups.

These findings show that the Colombian exam reform failed to reduce earnings inequality for multiple reasons. The new exam shifted some students with negative returns to college quality into higher ranked schools, but it also did a worse job matching students with high returns to top colleges. Thus, lowering test score gaps may not help low SES students if the new test is worse at identifying which students would benefit from attending a better college.

My paper adds to the literature on college mismatch by showing evidence of mismatch in the labor market. A few papers find evidence that affirmative action policies can, in some cases, lead to lower graduation rates for minority students (Arcidiacono et al., 2011, 2014, 2016).[1] I show that mismatch in graduation outcomes can also lower students' future earnings. Further, I show that mismatch can arise not just because some students have negative returns to college quality, but also because there is substantial heterogeneity in these returns that can be mis-measured by admission policies.

My paper also helps to reconcile the mismatch literature with other work on the returns to college selectivity (Dale and Krueger, 2002, 2014; Hoekstra, 2009; Saavedra, 2009; Ketel et al., 2012; Hastings et al., 2013; Goodman et al., 2014; Canaan and Mouganie, 2015; Kirkebøen et al., 2016; Chetty et al., 2017). Papers in this literature often find a positive earnings effects for students who are marginally admitted to a more selective college, and in some cases, these returns are even larger for disadvantaged students. If there is substantial heterogeneity in the returns to college quality, one may or may not find mismatch depending on the empirical strategy. Research designs in the selectivity literature measure returns for students who necessarily meet a college's admission standards. In the mismatch literature, as in my context, large-scale affirmative action policies led some disadvantaged students to attend colleges at which they would have otherwise been far below admission standards. This suggests that mismatch may arise only when disadvantaged students have vastly different academic preparation than their peers.

More broadly this paper relates to work on matching problems where there are potential complementarities between the two populations (Bhattacharya, 2009; Graham, 2011; Graham et al., 2014, 2016). In education this includes work on other admission tests like the

[1] Another strand of research finds no direct evidence of mismatch from statewide affirmative actions bans (Cortes, 2010; Backes, 2012; Hinrichs, 2012, 2014).

SAT (Rothstein, 2004), as well as other large-scale matching mechanisms like centralized assignment (Gale and Shapley, 1962; Abdulkadiroğlu et al., 2005), affirmative action (Bertrand et al., 2010; Antonovics and Backes, 2014; Bagde et al., 2016), and "percent plans" (Cullen et al., 2013; Black et al., 2015; Kapor, 2015).

The paper proceeds as follows. Section 2 develops a model that relates admission test scores to post-college earnings. Section 3 describes the Colombian admission exam reform and its effects on the SES test score gap. Section 4 shows how the reform affected the SES gap in college quality. Section 5 analyzes the earnings effects of the reform and estimates heterogeneity in the returns to college quality. Section 6 concludes.

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

This section develops a simple model that relates the socioeconomic gap in college admission test scores to the distribution of post-college earnings. I first describe student characteristics and show how the design of an admission test affects the socioeconomic test score gap. I then define the return to college quality—the key parameter that links the markets for college admissions and post-college labor. Lastly, I show that the relationship between test scores and earnings inequality depends on how well the exam measures heterogeneity in the returns to college quality.

2.1. **Students.**

The market for college admissions consists of a large population of I students with two characteristics of interest: socioeconomic status and ability. Let X_i denote individual i's socioeconomic status (SES), which reflects a dimension on which policymakers have a desire to promote equity (e.g., parental income). For simplicity, I define X_i to be a binary measure of high SES with $Pr[X_i = 0] = Pr[X_i = 1] = 0.5$.

I define individual i's ability as a K dimensional vector $A_i = \{a_{1i}, \ldots, a_{Ki}\}$, 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 vary in their correlations with SES and with students' potential outcomes.[2]

2.2. **College admission exam.**

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

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

[2] In labor economics it is common to use a test score as a measure of an individual's ability (e.g., Farber and Gibbons, 1996). This paper instead defines ability as a multi-dimensional vector of performance on *potential* exam questions. I take this approach to compare different exam designs.

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 or the student's health on exam day.

Test scores are arbitrarily scaled, so let $T_i = \frac{T_i^* - \bar{T}^*}{\sigma_T}$ be the normalized score, where $\bar{T}^*$ is the mean raw score and σ_T^2 is the raw score variance. If there are a large number of questions then we can consider T_i to be distributed approximately $N(0, 1)$. This reflects the common practice of testing agencies to normalize exam scores.

2.3. **Test score gaps and validity.** College admission tests are often accused of being unfair because they have large test score gaps between socioeconomic groups. Let $\bar{T}_x = E[T_i | X_i = x]$ denote the mean standardized test score for SES group $x \in \{0, 1\}$. Then the test score gap is given by

$$\bar{T}_1 - \bar{T}_0 = \frac{\sum_{k=1}^{K} w_k (\bar{a}_{k1} - \bar{a}_{k0})}{\sigma_T}, \tag{1}$$

where $\bar{a}_{kx}$ is the mean of the k^{th} ability for SES group x. $\bar{T}_1 - \bar{T}_0$ measures the difference between the average high and low SES test scores in standard deviation units.

The testing agency can reduce the test score gap by assigning more weight w_k to abilities that have lower mean differences between SES groups, $\bar{a}_{k1} - \bar{a}_{k0}$. An important question is whether reducing the test score gap also affects the exam's ability to predict student outcomes. One way to reduce the test score gap is simply to increase the relative importance of randomness in the exam. For example, suppose the testing agency increases weight on questions that nearly all students can answer (i.e., abilities for which $\bar{a}_{k1} \approx \bar{a}_{k0} \approx 1$). This redesign lowers the SES test score gap but also increases the fraction of the exam's total variance that comes from randomness.[3]

It is thus useful to compare exams not just on their test score gaps but also on their predictive power. To do this one can follow testing agencies' common method of validating exams by calculating correlations of test scores with measures of college success. Let Y_i be a measure of college success such as GPA or graduation rates, and let σ_Y^2 denote its variance. Then a measure of an exam's "raw validity" (Rothstein, 2004) is the correlation coefficient

[3] Assigning more weight to easy questions increases the exam's noise-to-signal ratio, σ_e^2/σ_T^2. To see this note that $\sigma_T^2 = \sum_{k=1}^{K} w_k^2 \operatorname{Var}(a_{ki}) + 2 \sum_{j=1}^{K-1} \sum_{k=j+1}^{K} w_j w_k \operatorname{Cov}(a_{ji}, a_{ki}) + \sigma_e^2$. Abilities k for which $Pr[a_{ki} \approx 1]$ have low variance, $\operatorname{Var}(a_{ki})$, and low covariance with other abilities j, $\operatorname{Cov}(a_{ji}, a_{ki})$. Thus σ_e^2 is a more important determinant of the total score in exams that place high weight w_k on easy questions. A similar logic applies to exams that place more weight on hard questions.

between Y_i and T_i, which I denote by $\rho_{Y,T}$:

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

There is a potential tradeoff between the SES test score gap and exam validity. It is easy to write an exam that is uncorrelated with SES but also has low predictive power. For example, the trivial exam in which $\bar{a}_{k1} = \bar{a}_{k0} = 1$ for all $w_k > 0$ has no test score gap but also zero validity. It is harder to identify abilities that are weakly correlated with SES but highly correlated with student outcomes.

Below I show that if one cares about earnings inequality, what matters is not an exam's GPA or graduation validity but rather its ability to predict the labor market impacts of college admission.

2.4. **Colleges.** The college market consists of a large population of C colleges indexed by c. I characterize colleges by a measure of their quality that I denote by Q_c. College quality is a complicated and multi-dimensional object, but in this paper I adopt a simple definition. I define a college's quality to be 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 c. This definition of college quality is widely used in research on the earnings impact of college selectivity (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 about college admissions to simplify the model. First, all students have the same preferences over colleges as denoted by the quality measure Q_c. Thus any college with a larger value of Q_c is strictly preferred to another college by all students. Second, I assume that colleges are perfectly selective, so that each college admits students with a single test score value. Under these assumptions college quality is given by

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

so that the quality of a student's college is equal to her own exam score. Since T_i is a normalized score, college quality is also measured in standard deviation units. Furthermore, Q_c is a fixed characteristic because colleges always enroll students with the same normalized scores regardless of the structure of the admission exam.

This model is a highly simplified version of college admissions that ignores heterogeneity in student and school preferences. However, it reflects the degree to which testing agencies

influence college admissions in practice. Testing agencies provide a ranking of students' propensity to succeed at top colleges irrespective of student and school preferences.

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

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

Equation (5) defines earnings as a function of an individual-specific intercept α_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 β_i, which I refer to as the *return to college quality*. β_i measures the percentage change in earnings from a one standard deviation increase the quality of the college a student attends. This follows the literature on the earnings impacts of college selectivity (e.g, Dale and Krueger, 2002), but my focus is on potential heterogeneity in these returns. The i subscript on this parameter allows for different returns across individuals. In this model β_i is a characteristic of a student that reflects her suitability for attending a higher quality as defined by labor market returns. Although earnings returns are likely to differ at each college, β_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.

2.6. **Test score gaps and earnings outcomes.** This section shows how college admission reforms that reduce the SES test score gap can affect labor market outcomes. For this it is useful to compute the average log earnings for SES group x, which I denote by $\bar{w}_x = E[w_{ic}|X_i = x]$. Taking the conditional average of the earnings equation (5) and using the college quality assumption (4) yields

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

where $\bar{\alpha}_x = E[\alpha_i|X_i = x]$ and $\bar{\beta}_x = E[\beta_i|X_i = x]$.[4]

Consider how an exam reform that raises low SES test scores, $\bar{T}_0$, affects low SES earnings, $\bar{w}_0$. Equation (6) decomposes the effect of such a reform into three terms. The first term, $\bar{\alpha}_0$, is the average of the low SES intercepts, α_i. This term measures the fixed component of individual skill that does not depend on the characteristics of the admission test.

The second term, $\bar{\beta}_0 \bar{T}_0$, captures the effect of the increase in test scores on the average student. This term is the product of the mean low SES test score, $\bar{T}_0$, and the mean low SES return to college quality, $\bar{\beta}_0$. If the average low SES student receives a large earnings benefit from attending a better college, $\bar{\beta}_0 \gg 0$, then reforms that raise low SES test scores

[4] The derivation of equation (6) uses $E[\beta_i Q_c|X_i = x] = E[\beta_i T_i|X_i = x] = \bar{\beta}_x \bar{T}_x + \text{Cov}(\beta_i, T_i|X_i = x)$.

are likely to raise average earnings. Conversely if $\bar{\beta}_0 \leq 0$, then such reforms are unlikely to benefit low SES students on the labor market.

The third term, $\text{Cov}(\beta_i, T_i|X_i = 0)$, highlights the importance of heterogeneity in the return to college quality. This term shows that average earnings also depend on the correlation of test scores T_i with the return to college quality β_i. Tests that can identify which students have large returns to college quality, $\text{Cov}(\beta_i, T_i|X_i = 0) \gg 0$, lead to higher average earnings for low SES students. Tests that fail to capture this heterogeneity yield lower earnings.

The third term of equation (6) also highlights how earnings depend on the validity of an admission exam (equation (2)). The usual measures of exam validity depend on the correlation of test scores with measures of college success such as GPA or graduation rates, Y_i, whereas the earnings implications of the exam depend on the correlation of test scores with the return to college quality, β_i. Thus exam validity matters for average earnings only if there is a strong correlation between Y_i and β_i. If GPA and graduation rates are weak predictors students' earnings benefits from attending a better college, maximizing exam validity has a limited effect on students' labor market outcomes.[5]

Furthermore, the specific abilities a_{ki} that are measured by the admission exam matter only to the extent that there is heterogeneity in the return to college quality. If there is little variation in β_i, the design of the exam is less important because $\text{Cov}(\beta_i, T_i|X_i = 0) \approx 0$ for any values of T_i. Any reform that reduces the test score gap will have a similar earnings impact, even if this occurs by increasing randomness in the exam.[6] When there is significant heterogeneity in β_i, it becomes more important that the test can identity which students will benefit from attending a better college. Exams that raise low SES test scores while still identifying abilities a_{ki} that are highly correlated with β_i will lead to higher earnings than exams that merely increased randomness.

To formalize the above intuition, I develop propositions that relate the SES test score gap to two labor market outcomes of potential interest to policymakers. First, a policymaker may care about the efficiency implications of the admission exam as measured by average earnings across all students in the market. Using equation (6), average earnings are simply given by $\bar{w} = (\bar{w}_1 + \bar{w}_0)/2$. Second, policymakers may also care about equity as measured by the earnings gap between high and low SES students. In the model this is given by $\bar{w}_1 - \bar{w}_0$.

The link between the test score gap and these earnings outcomes depends on the characteristics of the return to college quality, β_i. It is useful to consider three different cases of the population distribution of β_i. Below I state the propositions and describe the main intuition.

[5] Appendix A.2 formalizes this point by decomposing the β terms in equation (6) into terms that are correlated and uncorrelated with Y_i.

[6] $\text{Cov}(\beta_i, T_i|X_i = x)$ 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 given by the second term in equation (6), $\bar{\beta}_x \bar{T}_x$.

The full derivation of the propositions is in Appendix A.1, and Appendix A.3 gives a simple numerical example of the results.

(1) *Constant returns to college quality.* Consider first the simple case in which the return to college quality is positive and constant across individuals, i.e., $\beta_i = \beta > 0$ for all i. With constant returns, the third term in equation (6) drops out, $\text{Cov}(\beta_i, T_i|X_i = x) = 0$, and the average return to college quality, $\bar{\beta}_x$, is the same for all SES groups. Reforms that decrease the test score gap, $\bar{T}_1 - \bar{T}_0$, will therefore 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.

Propositions 1 summarizes the constant returns case.

Proposition 1 (**Constant returns**). *If the return to college quality, β_i, is positive and constant across individuals ($\beta_i = \beta > 0$), then reducing the SES test score gap lowers the earnings gap but has no effect on average earnings.*

(2) *Complementarity between SES and β_i.* Suppose now that β_i is heterogenous and that the average return to college quality is larger for high SES students, $\bar{\beta}_1 > \bar{\beta}_0$. This complementarity means that reducing the test score gap has a large negative earnings impact on the average high SES student ($\bar{\beta}_1\bar{T}_1 \ll 0$) and a smaller positive earnings effect on the average low SES student ($\bar{\beta}_0\bar{T}_0 > 0$). All else equal, reforms that lower the test score gap will reduce average earnings in the market. The only exception to this is if the new test becomes a better measure of β_i *within* SES groups, i.e., if $\text{Cov}(\beta_i, T_i|X_i = x)$ increases. Whether or not this is possible depends on the extent of within-SES heterogeneity in the return to college quality. If β_i is strongly related to SES but varies little within SES groups, lowering the test score gap will necessarily reduce average earnings.

Proposition 2 summarizes the case of a complementarity between SES and β_i.

Proposition 2 (**Complementarity**). *If the return to college quality, β_i, is larger for high SES students on average ($\bar{\beta}_1 > \bar{\beta}_0$), then reducing the SES test score gap lowers average earnings unless the new test is a better measure of β_i within SES groups, (i.e., $\text{Cov}(\beta_i, T_i|X_i = x)$ increases).*

(3) *Mismatch.* Proposition 2 describes a case in which reducing the SES test score gap can lower average earnings in the market. But there are also conditions on β_i in which reducing test score gaps can actually harm 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 (6) shows that, all else equal, increasing low SES test scores, $\bar{T}_0$, lowers average low SES earnings, $\bar{w}_0$. 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 states that some disadvantaged students may be better off attending lower-ranked schools because 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 their average if $\text{Cov}(\beta_i, T_i | X_i = 0)$ decreases. 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 high 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 low SES students students are not matched to colleges in a way that maximizes their average earnings.[7] Reforms that raise low SES test scores can lead to mismatch if the new test is a poor measure of those students who are likely to benefit the most at top colleges.

Thus reducing the test score gap need not translate into higher earnings for low SES students if there is mismatch. This is summarized in Proposition 3.

Proposition 3 (**Mismatch**). *Reducing the test score gap can lower average earnings for low SES students if:*

- *The average return to college quality is negative for low SES students ($\bar{\beta}_0 < 0$); or*
- *The correlation between test scores and the return to college quality decreases for low SES students (i.e., $\text{Cov}(\beta_i, T_i | X_i = 0)$ falls).*

In sum, the implications of college admission tests for students' labor market outcomes depend on the distribution of returns to college quality, β_i. If β_i varies little across students, then reforms that reduce the test score gap reduce earnings inequality without a significant efficiency cost. If β_i is strongly correlated with SES, there may be an equity/efficiency tradeoff in reducing the test score gap. Lastly, such reforms can harm low SES students if there substantial variation in β_i, or if some students have negative returns. In this case

[7] 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 6; if $\bar{\beta}_0$ is large and positive and $\text{Var}(\beta_i | X_i = 0)$ is small, then increasing $\bar{T}_0$ raises $\bar{w}_0$.

it is important for the admission exam to measure abilities that predict students' potential returns to college quality.[8]

I now turn to an empirical analysis that estimates heterogeneity in β_i.

3. Admission exam reform in Colombia

This section provides background on the Colombian college system, my data and variable definitions, and a reform of the national admission exam. I show that this reform led to a sharp decrease in test score gaps across various measures of socioeconomic status (SES). I also show that the reform reduced the exam's validity as measured by its ability to predict students' college outcomes.

3.1. **Institutional background and data.**

In this paper I use three administrative datasets that correspond to different parts of the Colombian higher education system.

The first dataset includes records from a national standardized exam called the ICFES, which Colombian students are required to take to apply to college.[9] The Colombian exam is analogous to the SAT in the U.S., but it is taken by nearly all high school graduates regardless of whether they plan to attend college. In this paper I use individual administrative records from the testing agency that cover 1998–2001 exam takers, which provide each student's test scores, high school, and characteristics including several measures of SES. I also use records from the testing agency that contain scores on a field-specific college *exit* exam for 2004–2011 test takers.

The second dataset includes enrollment and graduation records from the Ministry of Education that cover the near universe of colleges in Colombia. Colombia's higher education system consists of public and private institutions with varying selectivity and degree offerings. As in the U.S., admissions are decentralized; students apply to individual colleges and each institution controls its own selection criteria.[10] Nonetheless, the Ministry of Education tracks individual-level enrollment and graduation at nearly all colleges in the country. I use the Ministry's records on all enrollees from 1998–2012, which contain each student's institution, program of study, and dates of entry and exit.

[8] Propositions 1–3 do not cover all distributional possibilities. The return to college quality may be larger for low SES students, i.e., $\bar{\beta}_0 > \bar{\beta}_1$. In this case lowering test score gaps can reduce earnings inequality while also raising average market earnings. Consistent with this possibility, some work finds that disadvantaged students have larger returns to admission at selective colleges (e.g., Dale and Krueger, 2002; Saavedra, 2009).

[9] ICFES stands for Institute for the Promotion of Higher Education, the former acronym for the agency that administers the exam. The ICFES exam is now named Saber 11°.

[10] Unlike in the U.S., Colombian students apply not just to individual colleges but to college/major combinations. Below I show that my results are mainly driven by college effects, not major effects. The conceptual and empirical results in this paper go through if I use college/major pairs instead of colleges.

My last data source includes earnings records from the Ministry of Social Protection. These data contain monthly earnings for any college enrollee employed 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 enrollment and graduation outcomes for 1998–2001 exam takers and 2008–2012 formal sector earnings for college enrollees.[11]

3.2. **Variable definitions.** In this paper I use three different definitions of socioeconomic inequality using characteristics measured at the time students took the admission exam:

(1) Gaps between the top and bottom quartiles of the family income distribution;[12]
(2) Gaps between students whose mothers attended college and students with primary school (or less) educated mothers;
(3) Gaps between students who attended high and low ranked high schools.[13]

In some analyses I also report the gender gap in outcomes. Appendix A.4 provides details on the definitions of these variables.

My key outcome variables are admission exam scores, college quality, graduation rates, 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 students in the pre-reform cohorts and convert it to percentiles in the same manner.[14] 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.[15]

Table 1 display summary statistics for these variables and SES measures. Across the 1998–2001 cohorts I observe 1.6 million high school graduates who took the national admission exam. Most of my analyses focus on the roughly 600,000 students who enrolled in college. Columns (C)–(F) show large gaps in all outcomes between SES groups. For example, students in the top family income quartile score 26 percentile points higher on the math subject of the admission exam than students in the bottom quartile. Top quartile students also

[11] Appendix A.4 provides details on the coverage of each dataset and the merge process.
[12] I define these quartiles relative to the population of college enrollees, not all exam takers.
[13] The ICFES testing agency classifies high schools into seven categories based on their mean admission exam performance. The top three categories are "high" ranks and the bottom three are "low" ranks. I use a high school's pre-reform rank and hold this definition fixed across cohorts. High school choice in Colombia is primarily determined by geography and tuition, and thus high school rank is strongly related to SES.
[14] The college mean admission score also averages across all six exam subjects.
[15] I observe earnings in 2008–2012, so 10–11 years post-exam are the two years for which I observe earnings for all of the 1998–2001 cohorts. I compute average daily earnings by dividing total annual earnings by the number of formal employment days in the year, demeaning by year and exam cohort, and averaging across the two years.

TABLE 1. Summary statistics

	(A)	(B)	(C)	(D)	(E)	(F)
	Total # of students		1998–1999 cohort means			
Characteristic	High school graduates	College enrollees	Admit exam math score (percentile)	College quality (percentile)	Graduated from college	Daily earnings (2009 USD)
Top 25 family income	223,410	152,345	70.4	62.9	0.48	17.45
Bottom 25 family income	656,657	154,172	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
Male	741,973	285,393	57.0	53.2	0.34	14.60
Female	902,287	327,556	46.5	48.3	0.44	13.70
All students	1,644,260	612,949	51.2	50.6	0.39	14.13

Notes: Columns (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 cohorts, with column (C) including all exam takers and columns (D)–(F) including only college enrollees. See Appendix A.4 for more details on the sample and variable definitions.

attend colleges ranked 20 percentile points higher on average, are 14 percentage points more likely to graduate from college, and earn roughly 40 percent more in the labor market.

3.3. **The 2000 admission exam reform.** The Colombian national exam was first administered in 1968 with the aim of supporting college admissions. As the exam achieved widespread coverage in the 1980s, the government additionally began to use admission exam results to evaluate high schools.

In the mid 1990s, policymakers concluded that the exam was poorly designed for the dual objective of college admissions and high school accountability. There was a perception that the exam rewarded innate ability and rote memorization more than the capacity to apply material one had learned. Critics argued that it was a poor measure of high school value added, and that scores were biased in favor of high SES students.

To address these concerns the testing agency overhauled the admission exam, altering the type of questions, the tested subjects, and the scoring system. The goal was to develop an exam that tested "competencies" rather than "content," and in doing so to align the exam with the high school curriculum and the skills that predict college success. Appendix A.5 gives further details on the changes to the exam structure, and it provides sample questions from the pre- and post-reform tests.

The new exam debuted in 2000 after more than five years of psychometric research. The exam overhaul was widely publicized in the preceding months including substantial coverage

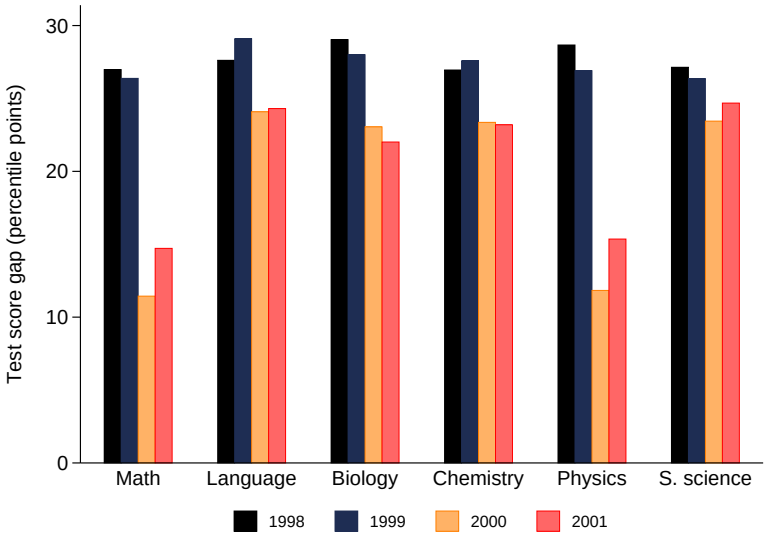


FIGURE 1. Family income test score gaps (Top 25 − Bottom 25)

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

in *El Tiempo*, the leading Colombia newspaper. In its objective and publicity, the redesign of the exam was thus similar to the overhaul of the U.S. SAT that debuted in 2016.

3.4. **Reduction in the SES test score gap.** Figure 1 shows that the 2000 Colombian admission exam reform led to a dramatic reduction in the test score gap between the top and bottom family income quartiles. The height of each bar corresponds to the difference in mean test percentiles between the top and bottom quartiles. In the pre-reform cohorts (1998–1999), the family income gap was approximately 27 percentile points in each of the six subjects of the exam. The reform had the largest effect on math and physics scores, with the SES test score gap falling by roughly 50 percent in the post-reform cohorts (2000–2001). The reform's effects were more modest in other subjects, yet in all cases the test score gap fell sharply by at least a few percentile points.

Table 2 shows that the reform also reduced test score gaps across other dimensions of SES. Column (A) displays the math test score gap in the pre-reform cohorts for each of my SES measures. Column (B) shows the change in the math test score gap in the post-reform cohorts.[16] The average difference in math test scores between students with college and primary educated mothers decreased dramatically, with a similar reduction in the test score gap defined by high school quality. The last row of Table 2 shows that the gender gap in math test scores also declined from roughly ten to three percentile points.

[16] For column (B), I regress math test scores on cohort dummies, an indicator for the high SES group, and the interaction of this indicator with a dummy for the post reform cohorts (2000–2001). Column (B) displays the coefficient on the interaction variable.

TABLE 2. Exam reform effects on test score gaps

	(A)	(B)	(C)	(D)
	Dependent variable: Math score (percentile)		Dependent variable: Mean score (percentile)	
	Pre-reform gap	Reform effect	Pre-reform gap	Reform effect
Family income gap	26.7	−13.7***	27.5	−8.8***
Top 25 − Bottom 25		(0.1)		(0.1)
Mother's education gap	23.4	−12.0***	24.1	−7.2***
College − Primary		(0.1)		(0.1)
High school rank gap	30.9	−17.1***	30.7	−9.9***
High − Low		(0.1)		(0.0)
Gender gap	9.9	−7.2***	7.7	−3.7***
Male − Female		(0.1)		(0.0)

Notes: Column (A) shows the math test score gap in percentile points for the 1998–1999 cohorts. For column (B), I regress math test scores on cohort dummies, an indicator for the high SES group, and the interaction of this indicator with a dummy for the 2000–2001 cohorts. Column (B) displays the coefficient on the interaction variable.

Columns (C)–(D) are analogous to the first two columns, but they report results from a stacked regression containing all six exam subjects. Each subject score in these regressions is weighted by the coefficients from a regression of college quality on the six exam scores in the 1998–1999 cohorts.

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

Columns (C)–(D) in Table 2 replicate the results in columns (A)–(B) using all exam subjects. These columns display coefficients from a regression that stacks all six exam subjects. Each subject score in this regression is weighted to reflect the fact that some subjects are more important for admission to top colleges than others. I estimate these weights by regressing college quality on the six exam scores in the pre-reform cohorts. This gives roughly twice as much weight to math scores as to chemistry, language, and social science scores, with biology and physics scores in between.[17] This stacked regression yields similar results as in columns (A)–(B). Across all SES measures and gender, the reform dramatically reduced test score gaps, with magnitudes of roughly 30 percent.

3.5. **Reduction in exam validity.** The results in Table 2 suggest that the testing agency achieved its goal of reducing socioeconomic bias in the exam. But a second objective of the reform was to design a better measure of the abilities that predict success in college. Did the new exam reduce test score gaps while still a remaining a good measure of college outcomes?

To address this question, I follow the standard practice of testing agencies and calculate measures of test validity. The most common definition of exam validity is the correlation

[17] These weights reflect the fact that one-third of students enroll in engineering programs. Results in columns (C)–(D) are similar using equal weighting, although the reform effects are slightly smaller in magnitude.

coefficient between test scores and measures of college success such as first year GPA or graduation rates (Section 2.3). To reduce the influence of a student's college choice, these calculations typically use residuals from regressions of both test scores and outcomes on college dummies (see e.g., Kobrin et al., 2008). Thus the correlations reflect only within college variation.

I use three measures of college success to measure the effects on exam validity. First, I use scores on a field-specific college *exit* exam that students take prior to graduation.[18] Second, I use an indicator for graduation from any college in my administrative records. Lastly, while I do not observe GPA in the administrative data, I have transcript records for some students in the 2000–2004 enrollment cohorts at one large flagship public university. For these students I calculate validity using their first year GPA.[19]

Table 3 shows that the reform generally reduced the validity of the admission exam. Column (A) shows the correlation between the exit exam score and each admission exam subject score (net of college dummies) in the pre-reform cohorts. Column (B) shows the change in these correlation coefficients in the post-reform cohorts.[20] The validity of the admission exam in predicting exit exam scores fell in all subjects except for chemistry, with declines of over 40 percent in math and physics—the two subjects with the largest reductions in the SES test score gap.

Columns (C)–(D) replicate these results using graduation rates as the measure of exam validity. The correlation between an indicator for college graduation and the math and physics admission scores declined by more than 50 percent with the exam reform. Graduation validity actually increased in the subjects for which the reform had a smaller effect on the SES test score gap, although the increases are modest.

Columns (E)–(F) also show declines in the validity of the math and physics exams using first year GPA from students at the flagship university. These calculations are underpowered because I observe fewer than 500 students who took the pre-reform admission exam.[21]

[18] The exit exam is now a national requirement for college graduation. During the period of my data it was optional, although many colleges required that their graduates take it. Roughly 40 percent of college enrollees in my sample took the exit exam, which reflects both its voluntary nature and that many students do not graduate. See MacLeod et al. (2017) for more details on the Colombian exit exam.

[19] See Appendix A.4 for details on the transcript data from this flagship university.

[20] Column (B) displays the coefficient on the interaction between the admission exam score and an indicator for the post reform cohorts in a regression that also includes cohort dummies and the admission exam scores. I normalize all variables to have standard deviation one within each cohort so that regression coefficients can be interpreted as correlation coefficients.

[21] Since my transcript records are for the 2000–2004 enrollment cohorts, I observe pre-reform scores only for students who waited one year or more after taking the admission exam to begin college. To address this issue, the regressions in columns (E)–(F) of Table 3 interact all covariates with indicators for the number of years since the admission exam. Thus correlations are calculated only from comparisons of students with the same number of years of delayed enrollment.

Table 3. Exam reform effects on test validity

	(A)	(B)	(C)	(D)	(E)	(F)
	Dependent variable: Exit exam score		Dependent variable: Graduated from college		Dependent variable: First year GPA	
	Pre-reform correlation	Reform effect	Pre-reform correlation	Reform effect	Pre-reform correlation	Reform effect
Math	0.369	-0.176^{***}	0.113	-0.058^{***}	0.154	-0.129^{**}
		(0.004)		(0.003)		(0.061)
Language	0.427	-0.088^{***}	0.105	0.009^{***}	0.059	-0.003
		(0.004)		(0.003)		(0.052)
Biology	0.388	-0.039^{***}	0.103	0.009^{***}	0.080	-0.050
		(0.004)		(0.003)		(0.057)
Chemistry	0.357	0.013^{***}	0.134	-0.003	0.132	-0.091
		(0.004)		(0.003)		(0.059)
Physics	0.359	-0.148^{***}	0.115	-0.063^{***}	0.073	-0.109^{*}
		(0.004)		(0.003)		(0.059)
Social science	0.418	-0.008^{**}	0.104	0.034^{***}	0.062	-0.018
		(0.004)		(0.003)		(0.063)
N	114,619	242,887	313,410	612,949	474	3,083

Notes: Admission exam scores and college outcomes in this table are residuals from regressing these variables on cells defined by a student's college and exam cohort. These variables are normalized to standard deviation one within each exam cohort so that coefficients can be interpreted as correlation coefficients.

Column (A) shows the correlation between the admission and exit exam scores in the 1998–1999 cohorts. For column (B), I regress the exit exam score on cohort dummies, the admission score, and the interaction of the admission score with a dummy for the 2000–2001 cohorts. Column (B) displays the coefficient on the interaction variable.

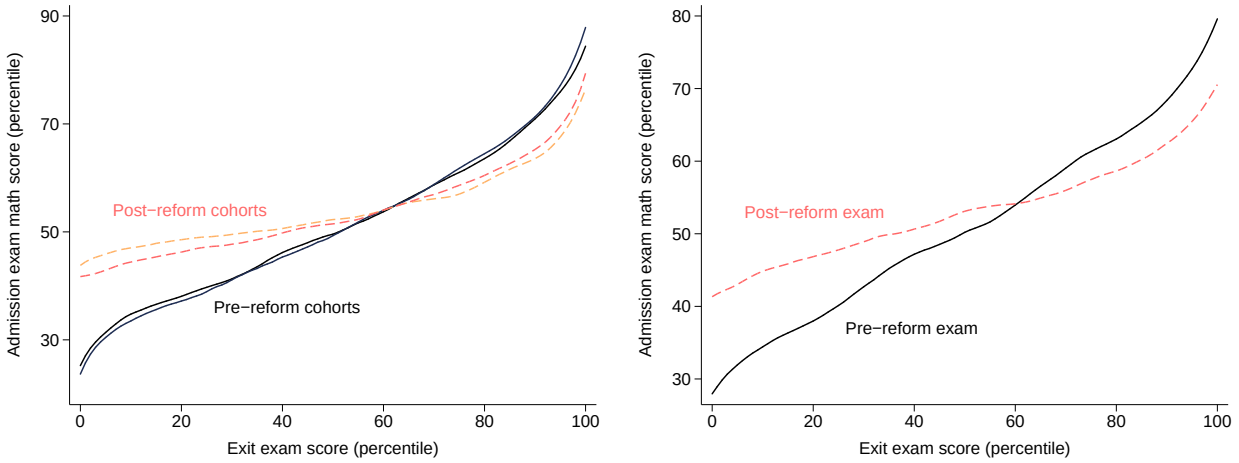
Columns (C)–(D) are analogous to columns (A)–(B) but with an indicator for college graduation as the dependent variable. Columns (E)–(F) are also analogous but with first year GPA as the dependent variable. The GPA regressions include only students in the 2000–2004 cohorts at a large public flagship university for whom I have transcript data. All covariates in these regressions are interacted with dummies for the number of years since the student took the admission exam.

Parentheses contain robust standard errors.

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

Nonetheless, the results suggest that the math and physics exams also became weaker predictors of first year GPA.

Table 3 shows that the reform reduced the admission exam's predictive power for students' college outcomes. Did this reduction in validity have a larger effect on high ability or low ability students? For example, one way to decrease exam validity is to convert challenging questions into medium-difficulty questions. This would affect the distribution of high and medium ability test scores, with few effects on low ability exam takers. Alternatively, a uniform reduction in exam difficulty would affect students of all ability levels.



Panel A. All exit exam takers

Panel B. Exit exam takers who took both admission exams

FIGURE 2. Math admission exam scores vs. exit exam scores

Notes: Panel A plots non-parametric regressions of the math percentile scores on the admission exam against percentile scores on the field-specific college exit exam. The sample includes all students who took the exit exam, with percentiles computed relative to each admission exam cohort from 1998–2001 within this sample ($N = 242{,}887$).

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

To answer this questions, one would ideally observe student performance on a separate exam taken earlier in high school. I do not have this information, but I can use exit exam performance as an alternate measure of ability. Exit exam scores are not a perfect measure of ability 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.

Figure 2 shows the distributional effects of the exam reform using exit exam scores as a measure of ability. The horizontal axis in Panel A is a student's percentile on the exit exam relative to all students who from her admission cohort who took both exams. The vertical axis shows the average percentile on the math component of the admission exam for each cohort, as estimated by a non-parametric regression. The dark solid lines show math scores for the pre-reform cohorts (1998 and 1999), and the lighter dashed lines plot math scores for the post-reform cohorts (2000 and 2001). There is a dramatic increase in admission math scores at the bottom of the exit exam distribution, and a large reduction in admission math scores at the top of the exit exam distribution. This shows that the reduction in validity occurred throughout the distribution of ability.

Panel B replicates the results in Panel A using only students who took *both* the pre-reform and the post-reform admission exams.[22] The two lines in this figure are thus computed using

[22] I observe roughly 10,000 students who took both admission exams as well as the exit exam.

the same sample, which helps alleviate the selection concerns described above. The pattern of results is similar to that using the full sample. Performance on the math component of the admission exam is much higher for students with low exit exam scores, and much lower for students with high exit exam scores. The admission exam reform led to a reduction in the exam's predictive power across the distribution of ability.

In sum, the Colombian admission exam reform reduced the SES test score gap, with dramatic declines in math and physics. However, the reduction test score gaps came at the cost of lower predictive power for students' college outcomes. The results are consistent with a new admission exam that was simply easier, and thus a larger proportion of the variance in test performance came from randomness. The question for the remainder of this paper is how the abilities that were measured (or mis-measured) by the admission exams relate to students' labor market outcomes.

4. Effects on the SES college quality gap

This section shows that the Colombian admission exam reform not only reduced test score gaps, it also reduced the SES gap in college quality. I first define a set of "treated" regions in which the reform was ex ante more likely to affect allocation of students to colleges because the national exam plays a larger role in admissions. I then develop a differences in differences specification exploiting variation in the reform's effects across regions and exam cohorts. Finally, I use this specification to show that the reform reduced the SES college quality gap, and I discuss several robustness tests.

4.1. **Definition of treated and control regions.**

I identify the exam reform's effects on students' college quality by exploiting two features of the Colombian higher education system that create regional variation in the stakes of the national exam.

First, most of Colombia's large administrative regions have a flagship public university that is a major player in the local market.[23] College attendance is highly regional, and in some cases the flagship school enrolls more than one-third of all area college students. Flagships are substantially less expensive than comparable private universities and give large tuition discounts to low-income students. Since financial aid markets are underdeveloped, the local flagship is often the only top college available to low SES students. As a result, flagship universities are the most selective colleges in the country. For example, the flagship university in Bogotá, Universidad Nacional de Colombia, typically admits less than ten percent of applicants, while the top private college, Universidad de Los Andes, has an admission rate near 50 percent.

[23] Colombia has 33 administrative *departamentos* that I call regions.



FIGURE 3. Definition of treated and control regions in Colombia

Notes: Dots represent flagship universities. Treated regions (in red) are those with flagships that use only the national exam scores for admissions (black dots). Control regions (in light yellow) are those with flagships that use other admission methods (white dots). Bogotá, which is its own administrative region, is a control region. The map does not show the island region of San Andrés y Providencia, which I define as a control region. I define treatment for regions with no universities using the closest flagship to their capital city. See Appendix A.6 for details.

The second relevant feature of the Colombia college system is that flagship universities control their own admission criteria. The majority of flagships admit students solely based on the national admission exam scores. These colleges compute major-specific weighted averages of the national exam subject scores and admit the highest ranking students up to a predetermined quota. Other flagships, however, require applicants to take the university's own exam, and some also consider high school grades or personal interviews. For example, Universidad Nacional de Colombia administers its own entrance exam, while Universidad del Valle, the flagship school in Cali, uses only national exam scores for admission.

I classify Colombia's 33 administrative regions as treated or control depending on the pre-reform admission method of their flagship university. For this I collected data on the pre-reform admission methods at each flagship university from historical websites and student regulations.[24] Black dots in Figure 3 represent cities with a flagship university that uses only national exam scores for admissions. White dots are cities in which the flagship uses its own entrance exam or other admissions criteria. Treated regions have flagships with national exam admissions and are shaded red in Figure 3. The light-colored control regions are

[24] Appendix A.6 provides details on the flagship universities, data sources, and admission methods.

TABLE 4. Summary statistics for pre-reform (1998–1999) college enrollees

	Treated regions	Control regions
Regions	22	11
Colleges in region	111	184
College enrollees/year	78,479	78,226
Enrolled in a public college	0.52	0.38
Enrolled in technical school/institute	0.22	0.37
Enrolled in region	0.58	0.84
Enrolled in any flagship university	0.34	0.17
Enrolled in national exam flagship	0.32	0.02

Notes: See Figure 3 and Appendix A.6 for the definitions of flagship universities and treated/control regions. The number of college enrollees per year is the average number calculated from the 1998–1999 cohorts.

those with other flagship admission methods. I define treatment for the sparsely-populated southeastern regions with no universities using the closest flagship to their capital city.[25]

Table 4 displays summary statistics for pre-reform college enrollees in the 22 treated and 11 control regions. Both areas have about 80,000 enrollees per year in my sample, but control regions have more colleges per student. Students in treated regions are more likely to attend public colleges, while control students are more likely to enroll in technical schools or institutes rather than universities.

The bottom of Table 4 shows how the features of the Colombian college system described above are useful for identification. Most students attend college in their own region, including 58 percent of treated students and 84 percent of control students. Flagship universities in both areas enroll a large fraction of all students, but there is substantial variation in exposure to flagships that use the national exam. Nearly one-third of treated students attend national exam admission flagships, compared with just two percent of control students.

The combination of regional markets and decentralized flagship admissions creates geographic variation in the potential impact of the national exam reform. The exam redesign was *ex ante* more likely to affect college admissions in treated regions, where flagship universities with national exam admissions comprise a large fraction of the market. The new exam altered the ranking of students in flagship admission pools and thus had the potential to alter the type of students received by many colleges in the region. The new ranking mattered less at flagship universities that used their own exams, so the reform was less likely to affect college admissions throughout control region markets.

[25] Bogotá is its own administrative region. Figure 3 does not display the island region of San Andrés y Providencia, which I define as a control region.

4.2. **Differences in differences specifications.** I use variation in the effects of the Colombian admission exam reform across treated and control regions to derive a benchmark differences in differences estimating equation.

First consider an equation that defines the test score gap,

$$T_{it} = \psi_t + \theta_t^T X_i + u_{it}, \tag{7}$$

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 3.2). Equation (7) is analogous to the definition of the test score gap in equation (1) from Section 2, but it estimates separate gaps for each exam cohort. The intercept ψ_t is thus 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 each cohort. The results in Section 3.4 show that test score gap, 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 region:

$$Q_c = \phi_r T_i + u_{icr}, \tag{8}$$

where Q_c measures college quality. As described in Section 3.2, I define Q_c to be a college's percentile rank based on the mean 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 ϕ_r thus gives the effect of individual test scores on college mean test scores in region r. The discussion in Section 4.1 implies that the ϕ_r coefficients are larger in treated regions.

Plugging equation (7) into equation (8) and renaming coefficients yields

$$Q_c = \gamma_{rt} + \theta_{rt} X_i + u_{icrt}. \tag{9}$$

The intercept in this regression, γ_{rt}, gives the average college quality for low SES students ($X_i = 0$) in region r and exam cohort t. The slope coefficient, θ_{rt}, gives the SES "college quality gap" in each region-cohort pair. The main prediction is that θ_{rt} should decline in treated regions relative to control regions 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_{rt}:

$$\theta_{rt} = \gamma_r + \gamma_t + \theta(\text{Treated}_r \times \text{Post}_t) + u_{rt}. \tag{10}$$

This regression includes region dummies, γ_r, and exam cohort dummies, γ_t. The variable of interest is the interaction of a dummy for treated regions, Treated_r, 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 regions relative to control regions.

Plugging equation (10) into equation (9) yields my main empirical specification:

$$Q_c = \gamma_{rt} + \Big(\gamma_r + \gamma_t + \theta(\text{Treated}_r \times \text{Post}_t)\Big)X_i + u_{icrt}. \tag{11}$$

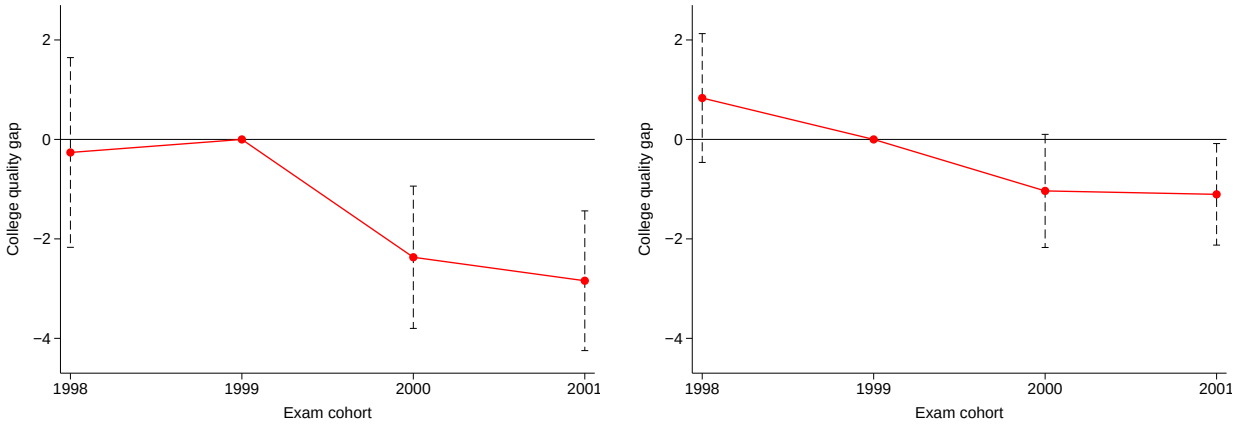
Equation (11) differs from a standard differences in differences regression in that it measures changes in a slope rather than changes in levels. In this case, the "slope" is the SES gap in college quality. The main prediction is $\theta < 0$, i.e., the college quality gap should decline more in treated regions than in control regions with the exam reform. In other words, in treated regions the reform should shift low SES students into higher quality colleges and displace high SES students to lower quality colleges.

The key identification assumption is the standard differences in differences assumption of parallel trends: the evolution of college quality, Q_c, would have been identical in treated and control regions in the absence of exam reform. To explore the validity of this assumption, I estimate versions of equation (11) that calculate separate coefficients for each exam cohort t. Plotting the θ_t coefficients provides visual tests for parallel trends in the pre-reform periods and sharp changes in outcomes following the exam reform.

4.3. **Effects on the SES college quality gap.** Figure 4 shows that the Colombian admission exam reform reduced the SES college quality gap in treated regions relative to control regions. The graph plot coefficients from event-study versions of equation (11), which calculate separate coefficients θ_t for each exam cohort t. Panels A–C correspond to the three SES gaps defined in Section 3.2, and Panel D depicts the gender gap. The horizontal axis in each panel shows the four exam cohorts, and the vertical axis displays the θ_t coefficients. The regression omits the interaction for the cohort right before the exam reform; this sets $\theta_{1999} = 0$ and makes all other coefficients relative to the 1999 cohort. Thus the plotted coefficients show how the difference in college quality gaps between treated and control regions varied across cohorts relative to 1999.

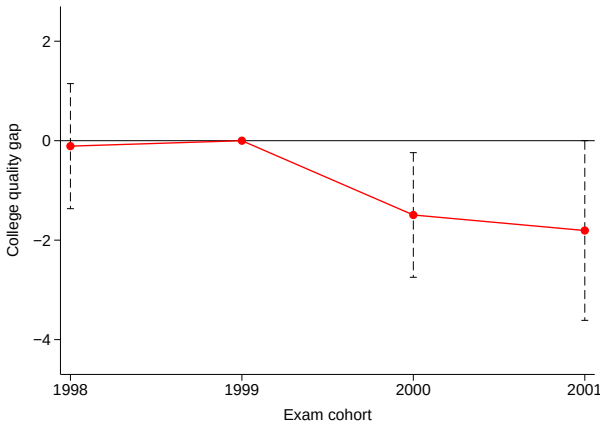
Panel A shows the effect of the reform on the college quality gap between the top and bottom family income quartiles. There is little evidence of differential pre-trends in the two cohorts before the reform; the average difference in the college quality gap between treated and control regions is almost the same in 1998 and 1999. In 2000, the first year of the new exam, the college quality gap falls sharply in treated regions relative to control regions. By the 2001 cohort, the college quality gap in family income is more than two percentile points lower in treated regions it was in 1999. In other words, treated region students with bottom quartile family incomes attended relatively higher quality colleges after the reform, while top-quartile students in treated regions were shifted into relatively lower quality colleges.

Panels B and C show similar results for college quality gaps as defined by mother's education and high school rank. For both SES measures, the college quality gap falls in treated

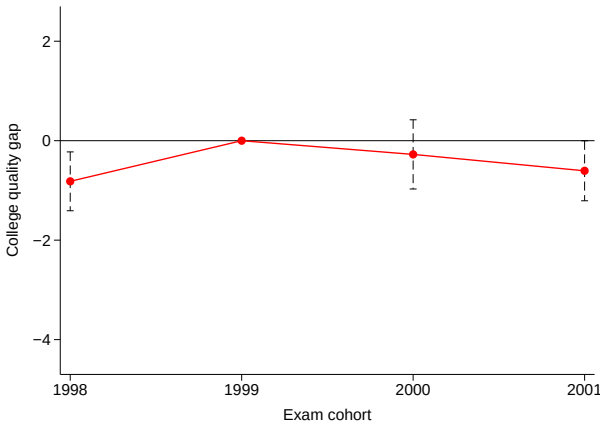


Panel A. Family income gap
Top 25 − Bottom 25

Panel B. Mother's education gap
College − Primary



Panel C. High school rank gap
High − Low



Panel D. Gender gap
Male − Female

FIGURE 4. Exam reform effects on the college quality gap

Notes: This figure plots coefficients θ_t from event study versions of equation (11), which interacts the θ coefficient with dummies for each cohorts t. The omitted interaction is between θ and the dummy for the 1999 exam cohort. The four panels correspond to the different SES and gender gaps defined in Section 3.2. Dashed lines are 90% confidence intervals using standard errors clustered at the region level.

regions relative to control regions in 2000, although for mother's education there is a pre-trend in the evolution of the college quality gap across regions. Panel D shows that the reform did not have a significant effect on the gender gap in college quality, despite the fact that the new admission test reduced the gender gap in test scores (Table 2).

Table 5 presents regression estimates analogous to Figure 4. Column (A) depicts the pre-reform college quality gap, which comes from estimating equation (9) for the 1998–1999 cohorts and removing the region and cohort interactions. Column (B) shows the estimates of θ from equation (11). The point estimates suggest that the reform reduced the family

TABLE 5. Exam reform effects on the college quality gap

	(A)	(B)	(C)	(D)
	Dependent variable: College quality (Q_c)		Dependent variable: Major quality	
	Pre-reform gap	Reform effect	Pre-reform gap	Reform effect
Family income gap	20.41	−2.47***	14.55	0.27
Top 25 − Bottom 25		(0.52)		(0.80)
Mother's education gap	17.04	−1.47***	12.09	0.45
College − Primary		(0.49)		(0.64)
High school rank gap	24.13	−1.59*	12.38	0.91
High − Low		(0.91)		(0.61)
Gender gap	4.92	−0.04	14.23	0.14
Male − Female		(0.24)		(0.32)

Notes: Column (A) depicts the pre-reform college quality gap for each measure of SES and gender, which comes from estimating equation (9) for the 1998–1999 cohorts and removing the region and cohort interactions. Column (B) shows the estimates of θ from equation (11).

Columns (C)–(D) are analogous to the first two columns, but the dependent variable is major quality rather than college quality. I use the Ministry of Education's classification of college programs into 54 field groups and define major quality as each field group's percentile rank based on the average admission score of its enrollees.

The sample for this table includes students in Column (B) of Table 1. Parentheses contain standard errors clustered at the region level.

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

income gap in college quality by 2.5 percentile points from a base of 20 percentile points. The magnitudes of the reform effects are similar for other SES measures. As in Panel D of Figure 4, the reform did not significantly affect on the gender gap in college quality.

4.4. **Alternative hypotheses.** This section considers other potential effects of the reform, and it argues that the primary effect was the reduction in the SES college quality gap.

Columns (C)–(D) in Table 5 provide evidence that the primary effect of the reform was on students' college quality rather than their major choices. Colombian students apply to college-major pairs, so it is possible that the exam reform also led to changes in the distribution of majors. The regressions in columns (C)–(D) use major quality, rather than college quality, as the dependent variable, defined as a major's percentile rank based on the average admission score of its enrollees.[26] Column (D) shows that the reform had little effect on the SES "major quality gap." This suggests that the primary effect of the exam reform was a change in the schools students attended rather than a change in their majors.[27]

[26] Majors are defined by the Ministry of Education's classification of college programs into 54 field groups.

[27] This result contrasts with the finding in Kirkebøen et al. (2016) that the earning returns to college-major pairs are primarily driven by major effects. One potential explanation for this result is a difference between the higher education markets in Norway—the setting for the Kirkebøen et al. (2016) paper—and Colombia.

Appendix Table A6 also shows that the results in column (B) of Table 5 are robust to the definition of treated and control regions. I use two alternative definition of treatment—one based on the closest flagship university to an individual's municipality, and the other based on the fraction of students in the pre-reform cohorts who enrolled in flagship universities that used the national exam for admissions. The main results are similar using either definition.

Appendix Table A7 shows results related to potential alternative effects of the admission reform. The reform had no detectable impact on the probability that high school graduates enrolled in college at all. Thus the primary effect of the exam reform was on *where* students went to college, not *whether* they went to college.[28] There is also little impact on students' likelihood of repeating the exam or delaying college enrollment, suggesting that the reform did not significantly alter students' exam-taking behavior. Lastly, I find that the reform did induce some low SES students to attend college closer to home. Thus the exam overhaul was not a perfect reallocation of college quality—colleges in treated regions saw modest increases in the size of their student bodies. Nonetheless, there are no significant changes in *average* college quality across regions, suggesting that students who switched regions were not systematically picking higher or lower quality colleges.

Taken together, these results suggest that the first order effect of the exam reform was an increase in college quality for low SES students in treated regions, and a relative decrease in college quality for high SES students. The next section examines how this reduction in the college quality gap affected students' labor market outcomes.

5. Labor market effects and the return to college quality

This section shows how the Colombian admission exam reform affected labor market outcomes, and it uses the reform to estimate heterogeneity in the returns to college quality. I first use a differences in differences specification to show that the reform lowered graduation rates and labor market earnings for both high and low SES students. I then develop an instrumental variables approach to estimate the average return to college quality for different SES groups. Next, I show how new machine learning tools (Athey et al., 2016) can be used to estimate heterogeneity in these returns within SES groups. Finally, I show how the

Anecdotally, the Colombian higher education system features greater variation in college reputation, while colleges in Norway differ less in their perceived rankings. The results in this paper also align with those in MacLeod et al. (2017), who find similar effects of the introduction of a Colombian college exit exam using college and college-major definitions of school reputation.

[28] This may be due to the fact that Colombian college markets have a large, open-enrollment sector where students can enroll if they are not admitted to top colleges. This also parallels the findings in research on other large-scale admission policies that primarily affect admission to selective colleges. For example, other work has found that affirmative action bans (Hinrichs, 2012) and percent plan admission rules (Daugherty et al., 2014) have little effect on the extensive margin of college enrollment.

reform affect the correlation of admission exam scores with the returns to college quality, and I discuss the results in the context of the propositions from Section 2.

5.1. **Effects on graduation rates and earnings.** To measure reduced form effects of the admission exam reform on outcomes, I use a standard differences in differences regression:

$$y_{irt} = \gamma_r^y + \gamma_t^y + \theta^y(\text{Treated}_r \times \text{Post}_t) + u_{irt}^y. \tag{12}$$

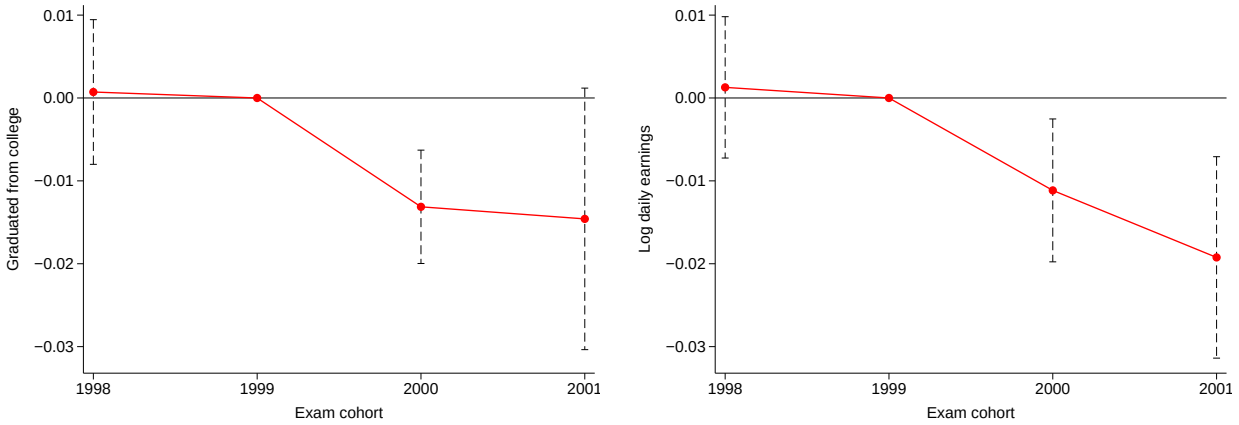
Equation (12) is analogous to specification (10), which estimated changes in the college quality gap, but it uses different dependent variables y_{irct}. This specification includes dummies for regions r and exam cohorts t. The variable of interest is the interaction between indicators for treated regions and post-reform cohorts. My main outcome is log earnings measured 10–11 years after the admission exam, and in this case the coefficient θ^y shows how the reform affected average earnings in treated regions relative to control regions. I also estimate equation (12) separately for different SES groups, which gives the change in average earnings for that SES group. In addition, I use an indicator for college graduation as the outcome variable y_{icrt} to explore the mechanisms underlying earnings effects.

Figure 5 shows that the admission exam reform decreased both average graduation rates and average post-college earnings in treated regions relative to control regions. The graphs display event-study coefficients θ_t^y from a version of equation (12) that interacts the coefficient of interest with cohort dummies, omitting the 1999 cohort. Panels A use college graduation as the outcome variable, and Panel B uses log daily earnings.[29] There is little evidence of a pre-trend in either outcome, and in both cases there is a sharp decline in the first post-reform cohort. The average graduation rate fell by about 1.5 percentage points in treated regions relative to control regions. Similarly, average daily earnings measured 10–11 years after the admission exam decreased by about 1.5 percent.[30] These results suggest that the reduction in the SES college quality gap led to an overall decline in both graduation rates and labor market earnings.

Columns (A)–(B) of Table 6 examine the source of these graduation and earnings declines. The last row shows the θ estimate from equation (12), which replicates the overall effects shown in Figure 5. The other rows present separate estimates of equation (12) for different SES groups. The main result is that the decline in earnings comes from both high SES students and low SES students. For example, earnings decreased by roughly two percent for students in the top quartile of the family income distribution, but also by roughly 1.5 percent for students in the bottom income quartile. This finding is similar across other definitions

[29] I observe earnings only for students who get jobs in the formal sector, but Appendix Table A7 shows that the reform had little effect on the probability of formal employment.

[30] Additional regressions suggest that the primary cause of the decrease in graduation rates was an increase in the probability of dropping out after the first year of college.



Panel A. Graduated from college

Panel B. Log daily earnings

FIGURE 5. Overall graduation and earnings effects

Notes: This figure plots event study coefficients θ_t^y from equation (12), with the interaction between $\text{Treated}_r \times \text{Post}_t$ and a dummy for the 1999 exam cohort as the omitted group. Dashed lines are 90% confidence intervals using standard errors clustered at the region level.

of SES. The pattern of results is also similar for graduation rates, although the graduation decline is more pronounced for high SES students.[31]

The negative effects for both high and low SES students are striking because these groups experienced opposite changes in average college quality. In particular, low SES graduation rates and earnings fell in treated regions despite the fact that they experienced increases in college quality on average. Furthermore, since the reduction in earnings was similar in magnitude for high and low SES students, the reduction in the college quality gap had essentially no impact on earnings inequality.[32]

5.2. **Instrumental variables specification.** Why did the reduction in the SES test score gap reduce low SES earnings? The framework in Section 2 showed that the relationship between test scores and earnings outcomes depends on the population distribution of the *return to college quality*. In this section I develop an instrumental variables (IV) approach to estimate heterogeneity in this parameter.

Recall that the individual return to college quality is the β_i coefficient from the earnings equation:

$$w_{ic} = \alpha_i + \beta_i Q_c + \epsilon_{ic}. \tag{13}$$

[31]Appendix Table A9 shows that the main results in Table 6 are robust to the wild t bootstrap procedure recommended by Cameron et al. (2008) for contexts with a relatively small number of clusters.

[32] Appendix Table A8 makes this point explicitly by estimating the reform's effects on SES earnings gaps. Because the negative graduation effects in Table 6 are larger for high SES students, the reform did lead to a reduction in the SES graduation gap. But this reduction came from high SES graduation rates falling more, not from increases in low SES graduation rates.

TABLE 6. Reduced form and 2SLS effects of the exam reform on outcomes

	(A)	(B)	(C)	(D)	(E)
	Reduced form effects		First stage and 2SLS		
$\text{Treated}_r \times \text{Post}_t \times \ldots$	Graduated from college	Log daily earnings	College quality/10	Return to college quality (IV)	First stage F statistic
Top 25 family income	-0.026^{***}	-0.022^{**}	-0.206^{***}	0.106^{*}	15.476
	(0.009)	(0.010)	(0.052)	(0.055)	
Bottom 25 family income	-0.009	-0.016^{***}	0.148^{*}	-0.111^{**}	3.402
	(0.008)	(0.005)	(0.080)	(0.052)	
College educated mother	-0.022^{**}	-0.014	-0.109^{*}	0.131	3.607
	(0.010)	(0.011)	(0.057)	(0.124)	
Primary educated mother	-0.005	-0.015^{***}	0.108	-0.138	1.757
	(0.007)	(0.005)	(0.082)	(0.099)	
High ranked high school	-0.020^{**}	-0.010	-0.082	0.118	1.256
	(0.007)	(0.009)	(0.074)	(0.132)	
Low ranked high school	-0.007	-0.015^{**}	0.122	-0.119	2.689
	(0.010)	(0.006)	(0.074)	(0.082)	
Male	-0.018^{**}	-0.018^{***}	0.008	-2.320	0.012
	(0.007)	(0.006)	(0.070)	(20.244)	
Female	-0.011	-0.012^{*}	0.031	-0.403	0.135
	(0.009)	(0.007)	(0.084)	(1.117)	
All students	-0.014^{*}	-0.016^{***}	0.018	-0.858	0.060
	(0.007)	(0.005)	(0.074)	(3.417)	

Notes: Columns (A)–(B) reports estimates of θ^y from equation (12) with graduation and log daily earnings as dependent variables. The last row shows the estimate of θ^y from the full sample, and the other rows show estimates for separate SES and gender groups.

Column (C) reports estimates of θ_x from equation (14). Column (D) reports estimates of β_x from the 2SLS system (14)–(15). Column (E) shows the first stage F statistics from equation (14).

The sample for the regressions in column (A) includes students in Column (B) of Table 1. The sample for the regressions in columns (B)–(D) include the subset of these students for whom I observe formal sector earnings (N = 340,623 overall). Parentheses contain standard errors clustered at the region level.

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

The challenge in estimating this parameter is that a student's choice of college quality, Q_c, is likely to be endogenous to her labor market outcomes. To address this endogeneity I modify the above differences in differences design and use only region and cohort variation to estimate this return. Consider a modified version of regression (12) with college quality, Q_c, as the dependent variable:

$$Q_c = \gamma_{rx} + \gamma_{tx} + \theta_x(\text{Treated}_r \times \text{Post}_t) + u_{icrtx}. \tag{14}$$

This specification differs from equation (12) in that it interacts all coefficients with dummies for SES groups defined by x. I begin by thinking of x as the SES groups defined in Section 3.2, but I show below that other methods of defining x are beneficial. Thus equation (14)

includes SES by region dummies, γ_{rx}, and SES by exam cohort dummies, γ_{tx}. The coefficient of interest, θ_x, measures the change in college quality for SES group x in treated regions relative to control regions.

I incorporate earnings by treating equation (14) as the first stage regression in a two stage least squares (2SLS) system. Adding the control variables from (14) into the earnings equation (13) gives the second stage regression:

$$w_{icrtx} = \pi_{rx} + \pi_{tx} + \beta_x Q_c + \epsilon_{icrtx}. \tag{15}$$

In this 2SLS system, the endogenous college quality variable Q_c is instrumented by the treatment variable ($\text{Treated}_r \times \text{Post}_t$) interacted with SES group dummies. β_x is thus an IV estimate of the return to college quality for SES group x.[33]

Column (C) in Table 6 shows estimates of θ_x from the first stage regression (14). I divide the dependent variable by ten so that one unit corresponds to ten percentile points in the distribution of college quality. Consistent with Table 5, college quality increased for the low SES group ($\theta_0 > 0$) and decreased for the high SES group ($\theta_1 < 0$).[34]

Column (D) shows estimates of β_x from the 2SLS system (14)–(15). These coefficients are equal to the Wald estimand that divides the reduced form earnings coefficients (column (B)) by the first stage coefficients (column (C)). Thus the results show a positive average return to college quality for high SES students, and a negative average return for low SES students. These returns are statistically significant for the family income groups, with a ten percentile increase in college quality leading to roughly a ten percent increase in earnings for top-quartile students and a ten percent decrease for low SES students. The returns are noisily estimated for other SES groups, and especially for men and women, an issue to which I now turn.

5.3. **IV assumptions and interpretation.** The 2SLS approach in columns (C)–(D) of Table 6 requires additional identification assumptions beyond those needed for the reduced form results in columns (A)–(B) (Angrist et al., 1996). One additional assumption is instrument relevance, which in this case requires that the exam reform altered the college quality of SES group x. In practice this means estimates of β_x are likely to be biased for SES groups whose college quality did not change substantially with the reform.

The importance of instrument relevance is illustrated by the first stage F statistics in column (E) of Table 6. The reform caused a large reduction in average college quality for students in the top quartile of family income, leading to a first stage F statistic of

[33] The β_x coefficients are equivalent to those from separate estimations of the 2SLS system (14)–(15) for each SES group.

[34] The results in Column (C) are equivalent to those in column (B) of Table 5, except the sample for the regression is restricted to those for whom I observe formal sector earnings.

15. In all other SES groups, however, F statistics are below the rule of thumb ($F = 10$) suggesting a weak first stage effect on that group's average college quality. This problem is particularly acute for the gender estimates, for which the first stage is especially weak. Thus the IV estimates of the return to college quality in column (D) should be interpreted with caution. Below I show how machine learning methods can help address the problem of weak instruments that arises from an arbitrary definition of SES groups.

In addition to relevance, IV also requires assumptions about instrument exclusion and monotonicity. The exclusion restriction in this context states that the exam reform did not affect students' earnings through channels other than changes in their college quality, Q_c. Monotonicity states that the exam reform shifted college quality in the same direction for all students in SES group x. For example, it assumes there are no low SES students whose value of Q_c decreased as a result of the exam reform.

Strictly speaking, the exclusion and monotonicity assumptions are unlikely to hold for two reasons. First, Q_c is a one-dimensional measure of college quality, while the returns to college choice may vary on multiple dimensions.[35] Q_c is also a fixed characteristic, meaning that it does not capture time varying aspects of college quality such as peer effects. Second, the effects of the reform on test scores are likely to vary within each SES group x. Monotonicity may be violated if, for example, some low SES students with high ability experienced *decreases* in test scores with the reform, and were thus shifted into lower quality colleges.

Despite the likely violation of these assumptions, estimates of β_x are useful summary statistics for how the benefits to attending a better school vary across students with different SES backgrounds. Since the goal of testing agencies is to produce a one-dimensional dimensional ranking of students, it is useful to know how this ranking maps into labor market outcomes using a related measure of college quality.

It is also important to note how the IV estimates, β_x, relate to the parameters in the model from Section 2. In the model, a key parameter is the average return to college quality for a certain SES group, which I denoted by $\bar{\beta}_x$ (equation (6)). Under the standard 2SLS assumptions, β_x is the average return to college quality *for those students in SES group x whose college quality was altered by the exam reform*—the "complier" population in the language of Angrist et al. (1996). In other words, $\bar{\beta}_x$ in the framework is the full population average for SES group x, while the 2SLS coefficient β_x is the complier population average.

While these parameters differ, Propositions 1–3 carry through if equation (6) is reinterpreted as defining complier average earnings rather than population average earnings. Thus the 2SLS estimates β_x are informative about the labor market implications of policies that produce similar complier populations as the Colombian admission exam reform.

[35] For example, public and private colleges with the same value of Q_c may have different support services for low SES students.

5.4. **Estimating heterogeneity in the returns to college quality.** Table 6 estimated the average returns to college quality for different SES groups, which relate to the parameters $\bar{\beta}_x$ in Section 2. But the framework also highlighted the importance the variation in these returns *within* SES groups. In particular, the main propositions in Section 2 relied on $\text{Cov}(\beta_i, T_i|X_i = x)$, the covariance between the returns to college quality and test scores conditional on SES. This section shows how new machine learning techniques (Athey et al., 2016) can be used to estimate within SES heterogeneity in these returns, while also helping with the instrument relevance problems discussed above.

The basic idea is to use additional covariates to estimate further heterogeneity in the returns to college quality. Let Z_i denote a vector of covariates that includes measures of SES as well as other individual characteristics such as gender, age, and number of siblings. Define groups z based on these covariates. One could then redefine the 2SLS system (14)–(15) with dummies for covariate groups z instead of SES groups x:

$$\begin{aligned} Q_c &= \gamma_{rz} + \gamma_{tz} + \theta_z(\text{Treated}_r \times \text{Post}_t) + u_{icrtz} \\ w_{icrtz} &= \pi_{rz} + \pi_{tz} + \beta_z Q_c + \epsilon_{icrtz}. \end{aligned}$$

In essence, this approach adds additional instruments for the endogenous college quality variable Q_c, which are interactions of the treatment variable ($\text{Treated}_r \times \text{Post}_t$) with dummies for individual characteristics. In theory, this method estimates a return to college quality β_z for each covariate group z. One could then analyze how the β_z coefficients vary within SES groups x to examine heterogeneity in the return to college quality.

There are, however, two problems with this approach. First, without direct knowledge of how the exam reform affected the test scores and college choices of different types of students, many of the covariate interactions with the treatment variable are likely to be weak instruments. This problem is evident from the first stage F statistics in Table 6, but the weak instrument problem becomes even more significant as one further partitions the sample population using covariates Z_i.

A second problem is that searching over covariates for heterogeneity in β_z can overstate the true amount of heterogeneity in this parameter. As Athey and Imbens (2016) point out, this is analogous to a researcher who reestimates the main estimating equation for different subsamples, and reports only estimates that yield statistically significant differences in the causal parameter of interest. This procedure creates upwardly biased estimates of the true heterogeneity in causal effects.

A solution to this problem comes from a recent set of papers that propose machine learning algorithms to estimate heterogeneity in causal parameters. Athey and Imbens (2016) show that regression tree algorithms can produce systematic estimates of heterogeneity in treatment effects that are not subject to subsample selection bias. Wager and Athey (2017)

extend this procedure from regression trees to random forests and prove consistency and asymptotic normality of such estimators. Lastly, Athey et al. (2016) extend the methods in Wager and Athey (2017) to apply not just to heterogeneity in treatment effects, but to heterogeneity in parameters from any set of local moment conditions, including instrumental variables regressions.

I use the random forest algorithms in Athey et al. (2016) to estimate heterogeneity in the returns to college quality across covariate groups z. To apply these methods, however, I must account for the fact that identification in my context relies on differences in differences variation across regions and exam cohorts. Define the variables $\tilde{w}_i$, $\tilde{Q}_c$, and $\widetilde{(\text{Treated}_r \times \text{Post}_t)}$ to be the residuals from regressions of log earnings, college quality, and treatment status on region dummies and cohort dummies. With these residuals one can define the 2SLS moment equations as

$$E\left[\left(\widetilde{\text{Treated}_r \times \text{Post}_t}\right)\left(\tilde{Q}_c - \theta(z)(\widetilde{\text{Treated}_r \times \text{Post}_t})\right)\right] = 0, \tag{16}$$

$$E\left[\left(\widetilde{\text{Treated}_r \times \text{Post}_t}\right)\left(\tilde{w}_i - \beta(z)\tilde{Q}_c\right)\right] = 0. \tag{17}$$

Equations (16)–(17) are standard IV moment conditions, but the causal parameters of interest, $\theta(z)$ and $\beta(z)$, can vary with other covariates, Z_i. In the authors' terminology, the algorithm that estimates $\theta(z)$ using moment condition (16) is a *causal forest* because it uses random forests to search for heterogeneity in the causal effect of being in a treated region and cohort on a student's college quality. The algorithm that estimates $\beta(z)$ using moment conditions (16)–(17) is an *instrumental forest* because it searches for heterogeneity in the causal effect of college quality on earnings using IV variation.

To implement the Athey et al. (2016) algorithms, I first define the covariates Z_i. Instead of specifying arbitrary SES groups as in Section 3.2, I allow the algorithm to search for optimal groupings of SES and other covariates. I include a large set of other covariates in Z_i: gender, age, number of siblings, family income, mother's and father's education, father's occupation, high school rank, high school academic type, and high school ownership status. I then use moment conditions (16)–(17) to estimate causal and instrumental forests on a random 50 percent subsample of my data—the "training" dataset.[36] This yields estimates of $\theta(z)$ and $\beta(z)$ for different covariate values $Z_i = z$. I predict values of these parameters into the other half of the data—the "validation" dataset—which I use for the results below.

Table 7 summarizes the first stage estimates of $\theta(z)$ from the causal forest algorithm. Column (A) displays the number of students in the validation data, for whom these statistics

[36] Specifically, I use the `causal_forest` and `instrumental_forest` functions from the authors' `grf` package for R.

TABLE 7. Causal forest first stage estimates, $\theta(z)$

	(A)	(B)	(C)	(D)	(E)	(F)
	N	Mean $\theta(z)$	St. dev. $\theta(z)$	% with $\theta(z) > 0$	Average SE	% with $F > 10$
Top 25 family income	42,220	-0.145	0.122	0.125	0.111	0.064
Bottom 25 family income	41,896	0.113	0.110	0.871	0.114	0.024
College educated mother	42,793	-0.095	0.155	0.296	0.110	0.060
Primary educated mother	69,358	0.102	0.114	0.828	0.109	0.034
High ranked high school	59,299	-0.052	0.160	0.389	0.108	0.052
Low ranked high school	51,344	0.097	0.118	0.807	0.108	0.034
Male	80,655	0.027	0.151	0.626	0.107	0.037
Female	89,657	0.041	0.154	0.637	0.108	0.044
All students	170,312	0.034	0.152	0.632	0.107	0.040

Notes: Column (A) shows the number of students in the 50 percent validation subsample. Column (B) reports the mean parameter value of $\theta(z)$ in this sample from estimating moment condition (16) using Athey et al. (2016)'s causal forest procedure on the training sample. I divide college quality, Q_c, by ten so that one unit corresponds to ten percentile points in the distribution of schools.

Columns (C)–(E) report the standard deviation of the $\theta(z)$ estimates, the fraction of these estimates that are positive, and the average standard error of these estimates. Column (F) shows the fraction of the $\theta(z)$ estimates with F statistics greater than 10.

are computed. Column (B) shows the mean of the $\theta(z)$ estimates for different SES groups and genders.[37] The mean first stage effects are similar to those reported in column (C) of Table 6, with low SES students experiencing increases in college quality on average, and high SES students experiencing average decreases in college quality. However, the causal forest procedure identifies significant heterogeneity in these effects within SES groups. As shown in column (C), within-group standard deviations of the first stage effects are greater than 0.1 (one percentile point of college quality). Furthermore, some student types experienced opposite changes in college quality relative to their peers in the same SES group. For example, column (D) shows that 13 percent of students in the bottom quartile of family income experienced decreases in college quality with the reform, while a similar fraction of top quartile students were shifted into higher quality schools.

Columns (E)–(F) of Table 7 describe the statistical power of the first stage estimates. Column (E) reports the average standard error of the $\theta(z)$ estimates as computed by the Athey et al. (2016) algorithm. Standard errors are on average larger than the magnitude of the parameter estimates, which suggests that most student types did not experience large changes in their college quality with the reform. As discussed above, the exam reform is a weak instrument for college quality for these student types, making estimates of their returns to college quality unreliable. Consistent with this, column (F) shows that a small fraction

[37] As in Table 6, I divide college quality by ten so that one unit equals ten percentile points.

TABLE 8. Instrumental forest estimates of the return to college quality, $\beta(z)$

	(A)	(B)	(C)	(D)	(E)	(F)
	N	Mean $\beta(z)$	St. dev. $\beta(z)$	% with $\beta(z) > 0$	Average SE	% with $p < 0.1$
Top 25 family income	2,685	0.051	0.073	0.727	0.092	0.106
Bottom 25 family income	1,003	-0.040	0.038	0.127	0.069	0.045
College educated mother	2,567	0.043	0.073	0.690	0.092	0.091
Primary educated mother	2,369	-0.039	0.053	0.173	0.082	0.035
High ranked high school	3,087	0.035	0.076	0.642	0.090	0.093
Low ranked high school	1,721	-0.044	0.055	0.151	0.082	0.038
Male	2,977	0.002	0.078	0.441	0.088	0.062
Female	3,920	-0.004	0.070	0.397	0.082	0.058
All students	6,897	-0.001	0.074	0.416	0.085	0.060

Notes: Column (A) shows the number of students in the 50 percent validation subsample who have first stage F statistics greater than 10 (column (F) in Table 7). Column (B) reports the mean parameter value of $\beta(z)$ in this sample from estimating moment conditions (16)–(17) using Athey et al. (2016)'s instrumental forest procedure on the training sample. I divide college quality, Q_c, by ten so that $\beta(z)$ can be interpreted as the earnings effect of a ten percentile point increase in college quality.

Columns (C)–(E) report the standard deviation of the $\beta(z)$ estimates, the fraction of these estimates that are positive, and the average standard error of these estimates. Column (F) shows the fraction of the $\beta(z)$ that are statistically different from zero at a 90% confidence level.

of students in each SES or gender group have estimates of $\theta(z)$ with an F statistic greater than the rule of thumb value of ten. For this reason, below I present estimates of the return to college quality for only student types with first stage F statistics greater than ten.

Table 8 summarizes estimates of the return to college quality, $\beta(z)$, from the instrumental forest estimation of moments (16)–(17). Column (A) shows the number of students for whom these statistics are computed, which equals the multiplication of columns (A) and (F) in Table 7. Column (B) shows the mean of the $\beta(z)$ estimates, which gives the average earnings effect of a ten percentile point increase in college quality. The mean estimates parallel those in column (D) of Table 6. High SES students have positive returns to college quality on average, and low SES students have negative average returns. The returns are generally smaller in magnitude than in Table 6. This is consistent with upward bias in the estimates when one ignores issues with instrument relevance.

The main takeaway from Table 8 is that there is significant heterogeneity in the returns to college quality both across and within SES groups. Column (C) shows that the standard deviation of the $\beta(z)$ estimates is roughly seven log points, and within-group standard deviations are generally of a similar magnitude. Further, column (D) shows that some low SES students have positive returns to college quality, and some high SES students have negative returns. This is illustrated in Figure 6, which plots the distribution of the $\beta(z)$ estimates for the top and bottom family income quartiles. The low SES distribution of returns is a

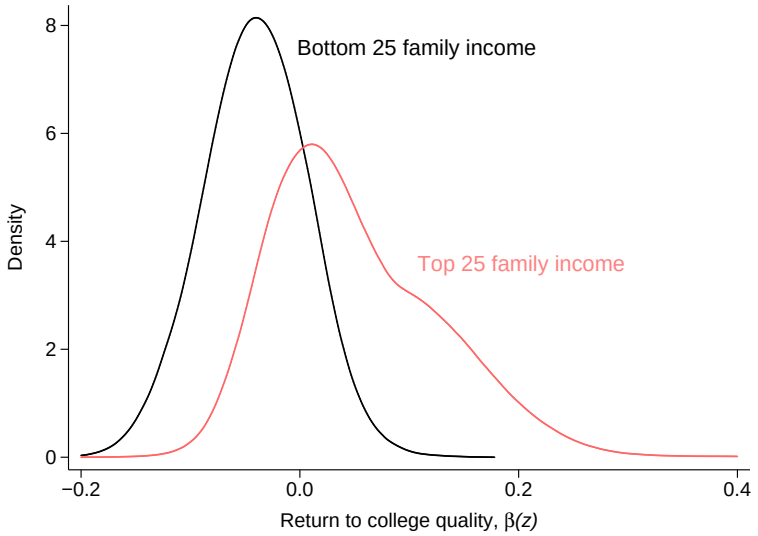


FIGURE 6. Returns to college quality, $\beta(z)$, for the top and bottom family income quartiles

Notes: This graph plots non-parametric densities of $\beta(z)$ for the top and bottom quartiles of family income. The sample includes students from the 50 percent validation subsample for whom the estimated first stage F statistic is greater than ten.

leftward shift of the high SES density, and the majority of the low SES $\beta(z)$ estimates are negative. Yet there is considerable heterogeneity in the returns to college quality within both family income groups.

Columns (E)–(F) of Table 8 display the average standard error of the $\beta(z)$ estimates and the fraction of these returns that are statistically significant at a 90% confidence level. High SES students are more likely to have statistically significant positive returns to college quality, although some low SES students have significant negative returns.

5.5. **Implications for Propositions 1–3.** The results in Table 8 show that there is scope for admission tests to measure (or mis-measure) heterogeneity in the returns to college quality. This last section describes how these results relate to the framework from Section 2 and to the observed earnings effects of the Colombian admission reform.

The framework contains three propositions on the implications of different distributions of the return to college quality, β_i. Proposition 1 states that if β_i is constant across individuals, a reduction in the SES test score gap has no effect on average earnings and lowers earnings inequality. This proposition is rejected in the Colombian context given the heterogeneity in returns shown in Table 8. The results are instead consistent with Proposition 2, which states that average earnings should decrease if there is a positive complementarity between SES and returns to college quality. The rightward shift of the SES distribution in Figure 6 is evidence of this complementarity, and it explains the negative earnings effects in Table 6 for both high SES students and for the full population.

TABLE 9. Reform effects on the correlation of returns to college quality, $\beta(z)$, and admission scores

	(A)	(B)	(C)	(D)
	All students		Bottom 25 family income	
Exam subject	Pre-reform correlation	Reform effect	Pre-reform correlation	Reform effect
Math	0.251***	-0.112***	0.094	-0.074
	(0.028)	(0.038)	(0.100)	(0.134)
Language	0.256***	-0.044	0.098	-0.109
	(0.028)	(0.038)	(0.098)	(0.133)
Biology	0.246***	0.000	0.045	-0.018
	(0.029)	(0.038)	(0.100)	(0.135)
Chemistry	0.264***	-0.017	0.062	-0.110
	(0.028)	(0.038)	(0.100)	(0.136)
Physics	0.259***	-0.067*	0.031	-0.010
	(0.028)	(0.038)	(0.100)	(0.136)
Social science	0.227***	-0.003	0.060	-0.149
	(0.028)	(0.038)	(0.099)	(0.135)
N	3,071	6,897	479	1,003

Notes: The sample includes students from the 50 percent validation subsample for whom the estimated first stage F statistic is greater than ten (column (A) in Table 8). Column (A) reports the correlation between admission exam scores and estimated returns to college quality, $\beta(z)$, in the 1998–1999 exam cohorts. Column (B) shows the difference in this correlation between the 2000–2001 cohorts and the 1998–1999 cohorts. Columns (C)–(D) show analogous results for the subsample of students with bottom quartile family incomes.

Parentheses contain robust standard errors adjusted for error in estimating $\beta(z)$.

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

Proposition 3 states conditions on the distribution of β_i under which reducing the test score gap can decrease average earnings for low SES students. Table 8 provides support for one of these conditions: the average return to college quality is negative for low SES students. Although most of the low SES estimates are not statistically different from zero, there are some low SES students with significant negative returns. This provides evidence that at least some low SES students in Colombia were mismatched at higher quality colleges.

Proposition 3 also describes another channel through which reducing test score gaps can lower earnings; earnings can fall if the new admission test becomes a worse measure of heterogeneity in the returns to college quality. To examine this possibility in the Colombian context, Table 9 shows how the reform affected the correlation of test scores with $\beta(z)$. Columns (A)–(B) present results for all students with first stage F statistics larger than ten. In the pre-reform cohorts, there is a positive correlation between admission scores and the returns to college quality on the order of about 0.25. This correlation falls in most exam subjects with the admission reform, with the largest declines in math and physics—the subjects with the biggest reductions in both the SES test score gap (Figure 1) and exam

validity (Table 3). Columns (C)–(D) show similar results for students with bottom quartile family incomes. The correlation coefficients are noisily estimated due to the small sample size, but in all subjects the reform reduced the admission exam's predictive power for low SES returns to college quality.

The results in Tables 8 and 9 suggest that the Colombian reform failed to reduce earnings inequality for multiple reasons. On the one hand, some low SES students with negative returns to college quality were shifted into higher ranked colleges. But the new admission exam also became a worse predictor of heterogeneity in the returns to college quality, reducing the rate at which students with positive returns were matched to top colleges. This shows that reducing the SES gap in college admission scores can potentially harm both high and low SES students if the new exam cannot identify which students would benefit from attending a better college.

6. Conclusion

A growing literature on college selectivity finds that, at least for certain individuals, attending a more selective college can lead to better career prospects (e.g., Dale and Krueger, 2002; Hoekstra, 2009). This work helps to explain why students and parents expend a great deal of energy on college admissions (Ramey and Ramey, 2010), as it addresses the salient question of whether college choice matters from an individual's perspective.

This paper has instead explored the consequences of college assignment from a market perspective. It asked how the matching of students to colleges via an admission test affects the distribution earnings in a college market. This follows in a long line of research on how the distribution of matches impacts aggregate outcomes (e.g., Gale and Shapley, 1962; Becker, 1973; Crawford and Knoer, 1981; Abdulkadiroğlu et al., 2005).

A market focus raises issues distinct from those that are relevant to a college-bound individual. One must examine the full distribution of students and colleges, not just students on the margin of admission at selective colleges. Further, there is the possibility of an equity/efficiency trade off in college admissions (Durlauf, 2008). For example, one can ask if the allocation of selective college slots to disadvantaged students raises or lowers the total productivity of a higher education system.

This paper used data and a natural experiment from Colombia to provide evidence on the market consequences of a large-scale reassignment of students to colleges. It analyzed a major reform of the national college admission exam that substantially reduced test score gaps between SES groups. In certain regions, the decline in test score gaps also reduced the SES gap in college quality. The primary effect of this reallocation of college quality was a market-wide reduction in graduation rates and post-college earnings. These negative effects

came from high SES students who were displaced to lower quality colleges, but also from low SES students who were shifted into higher quality colleges.

Combining these effects in an instrumental variables design, this paper documented significant heterogeneity in returns to college quality. In the Colombian context, these returns appear to be strongly correlated with SES and negative for some low SES students. This is consistent with a positive complementarity between socioeconomic status and college quality, and it shows that mismatch can arise in college assignment. Arcidiacono et al. (2016) find that science graduation rates would have been higher—for both minority students and for the University of California college system as a whole—in the absence of affirmative action policies. This paper extends these results to an entire college market and to labor market earnings after college.

An implication of these results is that one should be cautious in using estimates of the return to college quality for marginally admitted students to predict the effects of large-scale admission policies like affirmative action. Returns for students whom a college deems qualified may differ substantially from returns for students who would not be admitted under status quo admission procedures. This can explain why one might find large positive returns for low SES students who are on the margin of admission to top colleges (Hoekstra, 2009; Saavedra, 2009), and negative returns for students admitted under affirmative action (Arcidiacono et al., 2016).

Another takeaway is that reducing "bias" in college admission tests may not necessarily reduce inequality in labor market outcomes. Reforms that lower test score gaps—or eliminate the use of admission exams altogether—can potentially harm low SES students if there are no alternative ways of identifying which students would benefit from admission to top colleges. An important caveat, however, is that there may be other societal benefits of increased diversity (e.g., Rao, 2013; Fisman et al., 2016) that justify efforts to reduce the influence of socioeconomic background in college admissions. Furthermore, over time colleges may be able to adapt their teaching policies and support services to improve their students' outcomes.

References

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

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Antonovics, K. and B. Backes (2014). The effect of banning affirmative action on college admissions policies and student quality. *Journal of Human Resources 49*(2), 295–322.

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. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Athey, S., J. Tibshirani, and S. Wager (2016). Solving heterogeneous estimating equations with gradient forests. arXiv preprint arXiv:1610.01271.

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.

Becker, G. S. (1973). A theory of marriage: Part I. *Journal of Political Economy 81*(4), 813–846.

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.

Bhattacharya, D. (2009). Inferring optimal peer assignment from experimental data. *Journal of the American Statistical Association 104*(486), 486–500.

Black, S. E., K. E. Cortes, and J. A. Lincove (2015). Apply yourself: Racial and ethnic differences in college application. National Bureau of Economic Research.

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.

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.

Crawford, V. P. and E. M. Knoer (1981). Job matching with heterogeneous firms and workers. *Econometrica 49*(2), 437–450.

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.

Dale, S. B. and A. B. Krueger (2014). Estimating the effects of college characteristics over the career using administrative earnings data. *Journal of Human Resources 49*(2), 323–358.

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

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

Farber, H. S. and R. Gibbons (1996). Learning and wage dynamics. *The Quarterly Journal of Economics 111*(4), 1007–1047.

Fisman, R., D. Paravisini, and V. Vig (2016). Cultural proximity and loan outcomes. *American Economic Review* (forthcoming).

Gale, D. and L. S. Shapley (1962). College admissions and the stability of marriage. *The American Mathematical Monthly 69*(1), 9–15.

Goodman, J., M. Hurwitz, and J. Smith (2014). College access, initial college choice and degree completion. NBER Working Paper 20996.

Graham, B. (2011). Econometric methods for the analysis of assignment problems in the presence of complementarity and social spillovers. *Handbook of Social Economics 1*, 965–1052.

Graham, B. S., G. W. Imbens, and G. Ridder (2014). Complementarity and aggregate implications of assortative matching: A nonparametric analysis. *Quantitative Economics 5*(1), 29–66.

Graham, B. S., G. W. Imbens, and G. Ridder (2016). Identification and efficiency bounds for the average match function under conditionally exogenous matching. National Bureau of Economic Research Working Paper No. 22098.

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.

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

Kapor, A. (2015). Distributional effects of race-blind affirmative action. Working Paper.

Ketel, N., E. Leuven, H. Oosterbeek, and B. van der Klaauw (2012). The returns to medical school in a regulated labor market: Evidence from admission lotteries. 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.

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.

Ramey, G. and V. A. Ramey (2010). The rug rat race. *Brookings Papers on Economic Activity 41*(1 (Spring)), 129–199.

Rao, G. (2013). Familiarity does not breed contempt: Diversity, discrimination and generosity in Delhi Schools. Job Market Paper.

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.

Wager, S. and S. Athey (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* (just-accepted).

A. Appendix

A.1. **Test score gaps and earnings outcomes.** This section provides more details on the three propositions in Section 2, which describe how college admission reforms that reduce the SES test score gap can affect labor market outcomes.

I consider two labor market outcomes of potential interest to policymakers—one related to efficiency and one related to equity. A natural measure of efficiency is average log earnings in the market, $\bar{w} = E[w_{ic}]$, which is given by:[38]

$$\begin{aligned} \bar{w} &= \bar{\alpha} + E[\beta_i Q_c] \\ &= \bar{\alpha} + E[\beta_i T_i] \\ &= \bar{\alpha} + \text{Cov}(\beta_i, T_i) \qquad \text{(A1)} \\ &= \bar{\alpha} + \frac{(\bar{\beta}_1 - \bar{\beta}_0)(\bar{T}_1 - \bar{T}_0)}{4} + \sum_{x=0}^{1} \frac{\text{Cov}(\beta_i, T_i | X_i = x)}{2} \qquad \text{(A2)} \end{aligned}$$

where $\bar{\alpha} = E[\alpha_i]$ and $\bar{\beta}_1 - \bar{\beta}_0 = E[\beta_i | X_i = 1] - E[\beta_i | X_i = 0]$. Equations (A1) and (A2) are equivalent represents of mean earnings that are useful for the propositions 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:

$$\begin{aligned} \bar{w}_x &= \bar{\alpha}_x + E[\beta_i Q_c | X_i = x] \\ &= \bar{\alpha}_x + E[\beta_i T_i | X_i = x] \\ &= \bar{\alpha}_x + \bar{\beta}_x \bar{T}_x + \text{Cov}(\beta_i, T_i | X_i = x). \qquad \text{(A3)} \end{aligned}$$

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

$$\bar{w}_1 - \bar{w}_0 = \bar{\alpha}_1 - \bar{\alpha}_0 + \bar{\beta}(\bar{T}_1 - \bar{T}_0) + \text{Cov}(\beta_i, T_i | X_i = 1) - \text{Cov}(\beta_i, T_i | X_i = 0), \qquad \text{(A4)}$$

where $\bar{\beta} = E[\beta_i]$.[39] Equations (A3) and (A4) capture the implications of admission tests for SES-specific earnings and the degree of earnings inequality in the market.

Equations (A1)–(A4) show that the implications of exam reform for earnings equity and efficiency depend on the characteristics of the return to college quality, β_i. It is useful to discuss three different cases related to the distribution of β_i.

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

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

(1) *Constant returns to college quality.* Consider first the simple case in which the return to college quality is positive and constant across individuals, i.e., $\beta_i = \beta > 0$ for all i. Since β_i is fixed we have $\text{Cov}(\beta_i, T_i) = \text{Cov}(\beta_i, T_i|X_i = x) = 0$ in equations (A1) and (A4). Average earnings reduce to $\bar{w} = \bar{\alpha}$, and the SES earnings gap becomes $\bar{w}_1 - \bar{w}_0 = \bar{\alpha}_1 - \bar{\alpha}_0 + \beta(\bar{T}_1 - \bar{T}_0)$.

With a constant β_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.

Proposition 1 summarizes the constant return case.

Proposition 1 (**Constant returns**). *If the return to college quality, β_i, is positive and constant across individuals ($\beta_i = \beta > 0$), then reducing the SES test score gap lowers the earnings gap but has no effect on average earnings.*

(2) *Complementarity between SES and β_i.* Suppose now that β_i is heterogenous and that the average return to college quality is larger for high SES students, i.e., $\bar{\beta}_1 > \bar{\beta}_0$. Heterogeneity in β_i means that tests that assign higher scores to students with larger returns college quality will raise average earnings.[40]

If high SES students have larger returns to college quality on average, then average earnings are increasing in the test score gap since $\bar{\beta}_1 - \bar{\beta}_0 > 0$ in equation (A2). Exam reforms that reduce the test score gap will lower average earnings unless the new test becomes more related to β_i *within* SES groups, i.e., unless $\text{Cov}(\beta_i, T_i|X_i = x)$ increases. This covariance term depends on the types of questions in the exam, but it also depends on the within-SES variance in β_i. If β_i is strongly related to SES but varies little within SES groups, lowering the test score gap will necessarily reduce average earnings.[41]

Thus when there is a complementarity between SES and the return 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. This case is summarized in Proposition 2.

[40] For example, equation (A1) shows that average earnings are maximized by the test that maximizes $\text{Cov}(\beta_i, T_i)$.

[41] $\text{Cov}(\beta_i, T_i|X_i = x)$ 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 (A2)).

Proposition 2 (**Complementarity**). *If the return to college quality, β_i, is larger for high SES students on average ($\bar{\beta}_1 > \bar{\beta}_0$), then reducing the SES test score gap lowers average earnings unless the new test is a better measure of β_i within SES groups, (i.e., $\text{Cov}(\beta_i, T_i|X_i = x)$ increases).*

(3) *Mismatch.* Proposition 2 describes a case in which reducing the SES test score gap can lower average earnings in the market. But there are also conditions on β_i in which reducing test score gaps can actually harm 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 (A3) 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 disadvantaged students may be better off attending lower-ranked schools because 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 (A3)). 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 high 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 low SES students students are not matched to colleges in a way that maximizes their average earnings.[42] In this sense, reforms that raise low SES test scores can lead to mismatch if the new test is a poor measure of those students who are likely to benefit the most at top colleges.

Thus reducing the test score gap need not translate into higher earnings for low SES students if there is mismatch. This is summarized in Proposition 3.

Proposition 3 (**Mismatch**). *Reducing the test score gap can lower average earnings for low SES students if:*

- *The average return to college quality is negative for low SES students ($\bar{\beta}_0 < 0$); or*
- *The correlation between test scores and the return to college quality decreases for low SES students (i.e., $\text{Cov}(\beta_i, T_i|X_i = 0)$ falls).*

[42] 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 41; 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 (A3)).

A.2. **Exam validity and the return to college quality.** Section 2 shows that if the goal of a college admission exam is to promote efficiency and equity in earnings, the test should be a good measure of each student's return to college quality, β_i. The ideal in this sense is for the testing agency to assign high exam 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. Furthermore, β_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 β_i. In this last section, I compare the goal of measuring β_i to what testing agencies typically do in practice to evaluate their exams.

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

$$\beta_i = \pi + \phi Y_i + \tilde{\beta}_i.$$

Then from equation (A1), average earnings can be written as:

$$\begin{aligned} \bar{w} &= \bar{\alpha} + \text{Cov}(\pi + \phi Y_i + \tilde{\beta}_i, T_i) \\ &= \bar{\alpha} + \phi \text{Cov}(Y_i, T_i) + \text{Cov}(\tilde{\beta}_i, T_i) \\ &= \bar{\alpha} + \sigma_\beta \rho_{Y,\beta} \rho_{Y,T} + \text{Cov}(\tilde{\beta}_i, T_i) \end{aligned} \tag{A5}$$

where $\rho_{Y,\beta}$ is the correlation between Y_i and β_i, and σ_β^2 is the variance of β_i.

Equation (A5) shows that exam validity, $\rho_{Y,T}$, 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, β_i. If $\rho_{Y,\beta}$ is large and positive, then increasing exam validity raises average earnings. If $\rho_{Y,\beta}$ is small, then maximizing exam validity has a limited effect on students' labor market outcomes; average earnings is primarily determined by the part of β_i that is unrelated to Y_i.

Equation (A5) relates exam validity to average earnings, but an 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 (A3) shows that the relationship between differential validity and SES-specific earnings depends on the correlation of Y_i and β_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 β_i within SES groups.

TABLE A1. Test examples

Panel A. Student characteristics

Socioeconomic status, $X_i = \{0, 1\}$ with $Pr[X_i = 0] = Pr[X_i = 1] = 0.5$

Ability, $A_i = \{a_{1i}, a_{2i}, a_{3i}\} = \{X_i, a_i, 1\}$ where $a_i \sim N(0.2X_i, \sigma_a^2)$

Panel B. Test scores	(A) Maximum SES gap	(B) Measure of β_i^h	(C) Random test
Test weights, $w = \{w_1, w_2, w_3\}$	$\{1, 0, 0\}$	$\{0, 1, 0\}$	$\{0, 0, 1\}$
Raw test score, $T_i^* = A_i'w + e_i$	$X_i + e_i$	$a_i + e_i$	$1 + e_i$
High SES mean norm. score, $\bar{T}_1$	0.8	0.5	0
Low SES mean norm. score, $\bar{T}_0$	-0.8	-0.5	0
Test score gap, $\bar{T}_1 - \bar{T}_0$	1.6	1.0	0
Panel C. Constant return to college quality, $\beta_i^c = 0.1$			
$\text{Cov}(\beta_i^c, T_i \mid X_i = x)$	0	0	0
High SES mean earnings, $\bar{w}_1$	0.08	0.05	0
Low SES mean earnings, $\bar{w}_0$	-0.08	-0.05	0
Earnings gap, $\bar{w}_1 - \bar{w}_0$	0.16	0.10	0
Mean earnings, $\bar{w}$	0	0	0
Panel D. Heterogeneous returns to college quality, $\beta_i^h = a_i$			
$\text{Cov}(\beta_i^h, T_i \mid X_i = x)$	0	0.10	0
High SES mean earnings, $\bar{w}_1$	0.16	0.20	0
Low SES mean earnings, $\bar{w}_0$	0	0.10	0
Earnings gap, $\bar{w}_1 - \bar{w}_0$	0.16	0.10	0
Mean earnings, $\bar{w}$	0.08	0.15	0

Notes: The above calculations use $\alpha_i = 0$ for all i and thus $w_{ic} = \beta_i Q_c + \epsilon_{ic}$, where $Q_{c_i} = T_i$ and T_i is the normalized test score. I set $\text{Var}(e_i) = \sigma_e^2 = 0.01$ and $\sigma_a^2 = 0.02$ (equivalently, $\sigma_e = 0.10$ and $\sigma_a \approx 0.14$). Constant returns to college quality imply that $\bar{\beta}_1 = \bar{\beta}_0 = \bar{\beta} = 0.1$, while heterogenous returns imply $\bar{\beta}_1 = 0.2$, $\bar{\beta}_0 = 0$, and $\bar{\beta} = 0.1$.

A.3. **A simple example of test scores and returns to college quality.** Table A1 contains a simple example to illustrate the results in Propositions 1–3. In this example student ability is a three dimensional vector $A_i = \{a_{1i}, a_{2i}, a_{3i}\} = \{X_i, a_i, 1\}$. The first dimension of ability corresponds to an exam question that is perfectly correlated with SES, X_i. The second characteristic is given by $a_i \sim N(0.2X_i, \sigma_a^2)$, so a_i is partially correlated with SES with $E[a_i|X_i = 1] - E[a_i|X_i = 0] = 0.2$.[43] The third dimension of ability is $a_{3i} = 1$, which corresponds to an easy exam question for which all students know the answer.

Panel B of Table A1 describes three benchmark tests of these abilities. Column (A) is a test that assigns all weight to the first ability ($w_1 = 1$), so the raw test score is $T_i^* = X_i + e_i$. This

[43] Above I defined ability as a vector of binary indicators, a_{ki}. In Table A1 I assume $a_{2i} = a_i$ is normally distributed for simplicity, but this could be replicated by a vector of indicators for each value of a_i.

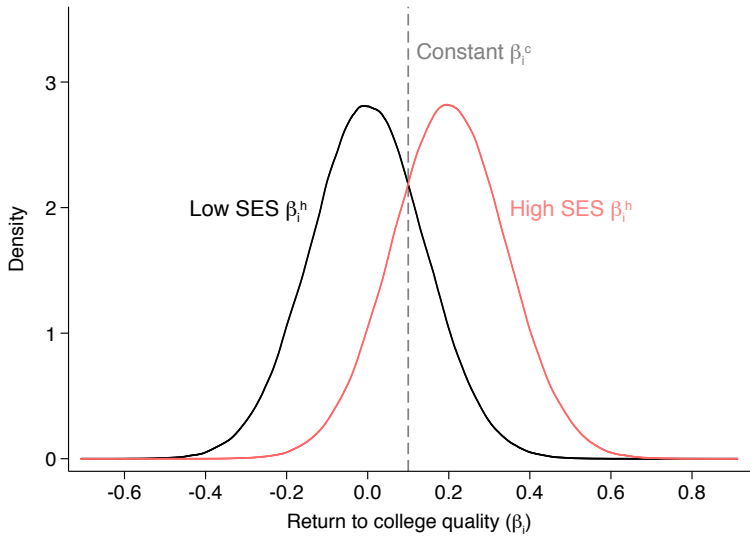


FIGURE A1. Example returns to college quality, β_i^c and β_i^h

Notes: This figure plots the example returns to college quality from Table A1. I assume $\beta_i^c = 0.1$ and $\beta_i^h = a_i$, where $a_i \sim N(0.2X_i, \sigma_a^2)$ with $\sigma_a^2 = 0.02$.

yields the maximum possible standardized SES test score gap of 1.6 standard deviations.[44] The test in column (B) assigns all weight to the second characteristic, so $T_i^* = a_i + e_i$. This reduces the test score gap to one standard deviation given the assumed values.[45] Lastly, the test in column (C) assigns all weight to the easy exam question, so $T_i^* = 1 + e_i$. All the variation in this test comes from randomness, so the test score gap falls to zero.

Panels C and D show the implications of these three tests for earnings given different assumptions about the return to college quality, β_i. Panel C assumes that the return to college quality is constant at $\beta_i^c = 0.1$ for all individuals i, as in Proposition 1. Assuming $\alpha_i = 0$ in the earnings equation (5), average earnings are $\bar{w} = \bar{\beta}^c \bar{Q}_c = 0.1\bar{T} = 0$. Switching between any of the three tests has no effect on average earnings. Average earnings for SES group x are $\bar{w}_x = 0.1\bar{T}_x$, so the earnings gap is directly proportional to the test score gap. Thus as the test score gap falls from 1.6 in test (A) to zero in test (C), earnings inequality, $\bar{w}_1 - \bar{w}_0$, falls from 0.16 to zero.

Panel D assumes that that the return to college quality is heterogenous with $\beta_i^h = a_i$. While β_i^h has the same mean as the constant β_i^c, it is positively correlated with SES with $\bar{\beta}_1^h = 0.2$ and $\bar{\beta}_0^h = 0$. In addition, β_i^h varies within SES groups with a standard deviation

[44] If the variance of e_i is small, then every high SES student scores above every low SES student on this test. Thus the average high SES test score is the mean of a standard normal variable truncated at zero, $\bar{T}_1 = E[T_i | T_i > 0] = \phi(0)/\big(1 - \Phi(0)\big) \approx 0.8$, and similarly $\bar{T}_0 \approx -0.8$.

[45] I assume $\sigma_a^2 = 0.02$ and $\sigma_e^2 = 0.01$ in Table A1. Thus $E[T_i | X_i = 1] = E[(T_i^* - \bar{T}^*)\sigma_T^{-1} | X_i = 1] = (0.2 - 0.1)\big((0.2)^2\sigma_X^2 + \sigma_a^2 + \sigma_e^2\big)^{-0.5} = 0.1(0.04 \times 0.25 + 0.02 + 0.01)^{-0.5} = 0.5$. Similarly, $E[T_i | X_i = 0] = -0.5$.

of $\sigma_a \approx 0.14$. Figure A1 depicts this distribution. The overlapping normal curves are the distribution of the heterogeneous β_i^h, while the vertical dashed line shows the constant β_i^c.

The test in column (A), $T_i^* = X_i + e_i$, captures the variation in β_i^h across but not within SES groups ($\text{Cov}(\beta_i^h, T_i | X_i = x) = 0$). This yields average earnings of $\bar{w} = 0.08$ and an earnings gap of $\bar{w}_1 - \bar{w}_0 = 0.16$. The test in column (B), $T_i^* = a_i + e_i$, is the best possible measure of the return to college quality given $\beta_i^h = a_i$, and thus it also captures the heterogeneity in β_i^h within SES groups ($\text{Cov}(\beta_i^h, T_i | X_i = x) = 0.1$). From equation (6), SES-specific earnings are $\bar{w}_1 = 0.2$ and $\bar{w}_1 = 0.1$. Average earnings under test (B) are therefore $\bar{w} = 0.15$. Lastly, the random test in column (C) misses heterogeneity in β_i^h altogether, and both mean earnings and earnings inequality are zero as before.

Panel D shows that when the return to college quality is heterogenous, earnings efficiency and equity no longer have a monotonic relationship with the SES test score gap. A reform that reduces the test score gap by switching from test (A) to test (C) lowers earnings inequality but also decreases average earnings. This illustrates the equity/efficiency tradeoff when there is a complementarity between β_i and SES, as in Proposition 2. Neither test captures the heterogeneity in β_i^h within SES groups, but test (C) also misses the heterogeneity across SES groups, lowering average earnings.

A reform that switches from test (B) to test (C) illustrates the mismatch possibility in Proposition 3. This reform causes low SES earnings to fall even as the test score gap shrinks. This arises because test (C) does a poor job at identifying which low SES students have high returns to college quality; students with negative values of β_i^h are shifted into better colleges. There is also an efficiency cost from the allocation of high SES students to lower quality colleges, causing average earnings to fall.

TABLE A2. Construction of analysis sample

	N
Total number of exam takers	2,100,424
Remove non 11^{th} graders	.
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.

A.4. **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 discusses how I select my main analysis sample for Sections 3–5. Finally, it describes the definitions of key variables.

The 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 whose high school identifier I observe in the data. In addition, I make a sample restriction that allows me to calculate family income quartiles for each student. I measure family income using income ranks in the spirit of Chetty et al. (2017). In the Colombian data, family income is grouped into ten bins based on multiples of the monthly minimum wage, but the distribution of these quartiles changes dramatically across cohorts due to variation in the Colombian inflation rate and minimum wage policy. To get a more stable measure of SES, I calculate predicted parental income using an individual's gender, mother's education, and high school. I then define family income quartiles based on a student's percentile rank of predicted parental income within their exam cohort, as described in Section 3.2.[46]

Table A2 shows the effect of these restrictions on sample size. The restrictions eliminate roughly ten percent of all exam takers in the ICFES records. The full sample includes roughly 1.6 million exam takers. However, most of my analyses are restricted to those who enrolled in college, as shown in the last row of Table A2. I test for and find no evidence of selection into college enrollment.

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

[46] Specifically, I define predicted parental income as the average 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 parental income.

TABLE A3. 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.

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 entrance exam. The exit exam became a requirement for graduation from any higher education institution in 2009.

Column (A) in Table A3 the 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 normative program duration.[47] 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 of exam takers depicted in column (B). Column (C) shows that the Ministry of Education records include all Universities but are missing a few colleges that provide more technical training.[48] Overall, 96 percent of exit exam takers attend colleges that appear in the Ministry of Education records.

I define my main measure of observable 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.[49] For the

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

[48] 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*).

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

20 colleges with fewer than ten pre-reform enrollees, I define Q_c as the mean exam score 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 average daily earnings by dividing monthly earnings by the number of formal employment days in each month and averaging across the year. In addition, I test for and find no evidence of selection into formal sector employment.

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 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 broadly comparable to the higher education enrollment rate in Colombia during the same time period.[50] 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. Among enrollees who took the admission exam between 1998 and 2001, I match 88 percent.[51]

A.5. **The 2000 admission exam reform.** This section provides further details on the 2000 reform of the Colombian college admission exam and the exam subjects that were offered between 1998 and 2006.

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.

[50] 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. Roughly 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.

[51] 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.

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.

Figures A2–A5 present sample questions from the biology, language, math, and social sciences components of the pre-reform and post-reform exams. The sample questions reinforce 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 typically 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 A4 shows the subject components that were included in the admission exam between 1998 and 2006. 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 3 I focus on the six subject groups listed in the leftmost column of Table A4: 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 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.

Table A4 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 approximately 50 in the pre-reform cohorts, and approximately 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'

Panel A. Pre-reform sample question

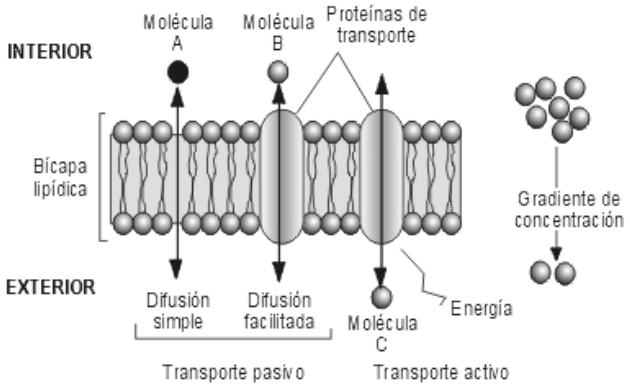
Which of the following two are inverse chemical processes?

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 A2. 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).

relative performance, and so for all my analysis I normalize exam scores to be mean zero and standard deviation one within each exam cohort.

A.6. **Flagship universities and their admission methods.** Table A5 shows the flagship universities in Colombia, and it describes how I define treated and control regions. Column

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 A3. 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

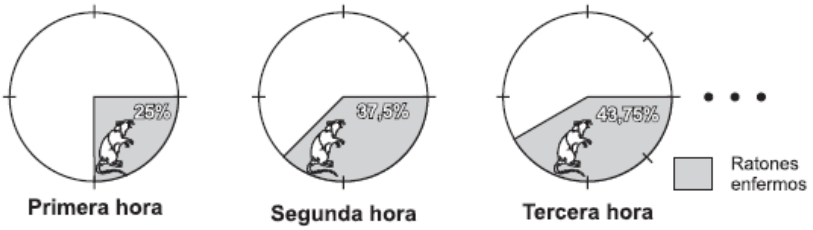
It is known that the result of multiplying a number by itself several times is 256. You can identify this number if it is known

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 A4. 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).

Panel A. Pre-reform sample question

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

Reason: 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 A5. 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).

TABLE A4. Mean admission score by exam component and cohort

Subject groups	Exam components	Exam cohort 1998	1999	2000	2001	2002	2003	2004	2005	2006
Biology	Biology	47.6	48.1	45.1	44.6	45.0	45.3	46.0	47.3	47.0
Chemistry	Chemistry	46.1	51.0	45.0	45.2	44.3	43.5	42.3	43.6	45.2
Language	Language	48.6	50.9	46.5	46.4	48.3	48.9	52.4	46.4	48.4
Math	Math aptitude	49.1	50.5							
	Math knowledge	49.0	49.0							
	Math			43.0	41.1	42.7	41.8	41.0	44.5	45.7
Physics	Physics	47.3	47.1	45.3	46.7	45.3	46.2	42.9	46.7	45.9
S. sciences	Social sciences	47.9	48.6							44.9
	Geography			44.4	43.0	43.4	42.9	49.6	41.3	
	History			43.5	43.5	43.4	43.3	44.2	42.5	
Excluded components	Verbal aptitude	48.3								
	Philosophy			44.8	43.9	44.7	44.9	45.4	43.6	47.1
	Foreign language			41.0	42.3	42.1	41.7	39.6	43.4	43.1
	Elective	49.5	50.9	52.3	55.4	52.4	48.5	48.7	47.5	48.0
	Mean (all components)	48.2	49.5	45.1	45.2	45.2	44.7	45.2	44.7	46.2
	St. dev. (all components)	10.2	10.2	7.4	7.5	7.4	7.4	8.0	8.0	8.0

Notes: The sample includes only students who took the exam in the year of their high school graduation.

TABLE A5. Flagship universities and definition of treated regions

(A) Region	(B) No. of exam takers	(C) College name	(D) City	(E) Type	(F) Admission method	(G) Closest other region	(H) Treated region?
Bogota	79,592	Universidad Nacional de Colombia	Bogota D.C.	University	Other		
Antioquia	56,374	Universidad de Antioquia	Medellin	University	Other		
Valle	49,166	Universidad del Valle	Cali	University	National exam		✓
Atlantico	25,471	Universidad del Atlantico	Barranquilla	University	Other		
Santander	23,614	Universidad Industrial de Santander	Bucaramanga	University	National exam		✓
Cundinamarca	21,193	Universidad de Cundinamarca	Fusagasuga	University	National exam		✓
Bolivar	16,356	Universidad de Cartagena	Cartagena	University	Other		
Boyaca	16,098	Universidad Pedagogica y Tecnologica de Colombia	Tunja	University	National exam		✓
Cordoba	15,237	Universidad de Cordoba	Monteria	University	Other		
Tolima	14,454	Universidad del Tolima	Ibague	University	National exam		✓
Narino	13,725	Universidad de Narino	Pasto	University	National exam		✓
Cauca	12,662	Universidad del Cauca	Popayan	University	National exam		✓
Norte Santander	12,454	Universidad Francisco de Paula Santander	Cucuta	University	National exam		✓
Caldas	11,172	Universidad de Caldas	Manizales	University	National exam		✓
Huila	10,688	Universidad Surcolombiana	Neiva	University	National exam		✓
Magdalena	10,115	Universidad del Magdalena	Santa Marta	University	Other		
Risaralda	9,988	Universidad Tecnologica de Pereira	Pereira	University	National exam		✓
Cesar	8,374	Universidad Popular del Cesar	Valledupar	University	National exam		✓
Meta	7,790	Universidad de Los Llanos	Villavicencio	University	National exam		✓
Sucre	6,751	Universidad de Sucre	Sincelejo	University	National exam		✓
Quindio	6,319	Universidad del Quindio	Armenia	University	National exam		✓
La Guajira	4,932	Universidad de la Guajira	Riohacha	University	Other		
Choco	3,050	Universidad Tecnologica del Chocodiego Luis Cordoba	Quibdo	University	Other		
Caqueta	2,545	Universidad de la Amazonia	Florencia	University	National exam		✓
Casanare	2,004	Fundacion Univ. Internacional del Tropico Americano	Yopal	Univ. Inst.			
Arauca	1,669					Norte Santander	✓
Putumayo	1,578	Inst. Tecnologico del Putumayo	Mocoa	Tech. Inst.			
San Andres	664	Inst. Nacional de Formacion Tech. Prof. de San Andres	San Andres	Tech. Inst.			
Amazonas	512					Caqueta	✓
Guaviare	250					Meta	✓
Vichada	122					Norte Santander	✓
Vaupes	78					Meta	✓
Guainia	74					Meta	✓

Notes: The number of exam takers is the average number per year calculated from the 1998–1999 cohorts.

(A) lists the 33 administrative departments in Colombia that I call regions, and column (B) shows the average number of national exam takers per year in the 1998–1999 cohorts.

I define flagships as the largest public university in each region. "Largest" is defined by the total number of students for all cohorts in my enrollment records, and the restriction to "universities" excludes colleges that the Ministry of Education classifies as "university institutes" or "technical institutes." In addition, I make two modification to this definition of flagships. First, I consider Universidad Nacional de Colombia—which is the most prestigious public college in the country—to be the flagship school in Bogotá even though there are several other public universities with more enrollees.[52] Second, in the region of Norte Santander, I consider Universidad Francisco de Paula Santander to be the flagship since it is located in the capital city of Cúcuta, even though the public university in the city of Pamplona has more enrollees.[53] Column (C) lists the flagship universities in each region under this definition, and column (D) shows the cities in which they are located. In three regions, the largest public college is not a university (see column (E)), so I classify these regions as not treated. This classification captures the fact that non-university level colleges are typically open enrollment, and thus the exam reform is not likely to affect admission to these colleges.

Column (F) lists the admission method that each flagship used before the 2000 reform. I collected information on pre-reform admission methods 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. The majority of flagships used only national exam scores for admission. Eight flagships used other admission methods, which most commonly meant that they required applicants to take the university's own admission exam. In some cases these flagships also considered other information such as high school GPA or personal interviews.

Six regions of Colombia do not contain any colleges in my records. To classify treatment, I assign these regions to the closest region with a flagship using the distance between capital cities. Column (G) shows the closest assigned region for these six regions without any colleges.

Finally, column (H) of Table A5 shows my classification of regions as treated or control. Treated regions are those with flagship universities that use only the national exam for admissions, or those regions without colleges whose closest region has a national exam flagship. Control regions are those with flagship universities that use other admission methods,

[52] Bogotá is the capital of Colombia and is its own administrative region.

[53] This does not affect the classification of Norte Sandander as a treated region as both colleges used only national exam scores for admissions.

or those regions whose largest college is not a flagship. Summary statistics on the college markets and student populations in treated and control regions are in Table 4.

A.7. **Alternative hypotheses.** This section considers other effects of the exam overhaul, and it argues that these alternative hypotheses are less likely to explain the negative graduation and earnings effects than the changes in the distribution of college quality.

Table A6 examines the robustness of the result in Table 5 that the reform reduces the SES college quality gap. Columns (A)-(D) report results from regressions analogous to columns (A)–(B) in Table 5, but using different definitions of college quality. Columns (A)–(B) define Q_c using the average pre-reform admission exam score at the college-major level rather than the college level. This reflects the fact that in Colombia admissions are to college-major pairs. Thus it is possible that the exam reform led to changes in the distribution of students' majors in addition to changes in the distribution of their colleges. Columns (A)–(B) show that the effects of the reform on the SES college quality gap are similar using the college and college-major definitions of quality. Columns (C)–(D) repeat the results from Columns (C)–(D) in Table 5, which shows that when I define Q_c at the major level only, the estimated effect of the reform is close to zero. This suggests that the primary effect of the exam reform was a change in the schools students attended rather than a change in their majors.[54]

[54] This result contrasts with the finding in Kirkebøen et al. (2016) that the earning returns to college-major pairs are primarily driven by major effects. One potential explanation for this result is a difference between the higher education markets in Norway—the setting for the Kirkebøen et al. (2016) paper—and Colombia. Anecdotally, the Colombian higher education system features greater variation in college reputation, while colleges in Norway differ less in their perceived rankings. The results in this paper also align with those in MacLeod et al. (2017), who find similar effects of the introduction of a Colombian college exit exam using college and college-major definitions of school reputation.

TABLE A6. Exam reform effects on the college quality gap

	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
	Q_c defined at college-major level		Q_c defined at major level only		Treatment_m defined by closest flagship to municipality		Treatment_m defined by pre-reform enrollment in municipality	
	Pre-reform gap	Reform effect	Pre-reform gap	Reform effect	Pre-reform gap	Reform effect	Pre-reform gap	Reform effect
Family income gap	24.41	−1.88***	14.54	0.27	20.38	−3.11***	20.38	−3.24***
Top 25 − Bottom 25		(0.67)		(0.80)		(0.61)		(0.60)
Mother's education gap	19.57	−0.68	11.94	0.45	16.55	−1.67***	16.55	−1.66***
College − Primary		(0.53)		(0.64)		(0.52)		(0.58)
High school rank gap	23.72	−0.81	12.27	0.91	20.44	−2.09**	20.44	−2.79***
High − Low		(0.87)		(0.61)		(1.02)		(0.91)
Gender gap	10.21	−0.10	14.26	0.14	5.03	−0.12	5.03	−0.34
Male − Female		(0.45)		(0.32)		(0.24)		(0.27)

Notes: The table reports estimates of θ from equation (11). Regressions in columns (D) and (E) replace region fixed effects with municipality fixed effects. See the text for descriptions of the variables in each column. Parentheses contain standard errors clustered at the region level. The pre-reform mean college quality gap is calculated from the 1998–1999 cohorts.

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

Columns (E)–(H) show that these first stage effects are robust to different definitions of treated and control regions. The benchmark specification defines the treatment variable, Treated_r, by whether or not the flagship university in the region where a student attended high school uses only the national exam score for admissions. Columns (E)–(F) defines treatment instead using the closest flagship university to a students' municipality. This reflects the fact some students are likely to attend college out-of-region if their high school is close to a major city in a different region. The reform's effects on the college quality gap are similar using this more granular definition of treatment.[55] Column (E) also uses a municipality-level definition of Treated_r, but it is even more granular in that it allows for different intensities of treatment. Treatment is defined using the fraction of pre-reform college students from each municipality who enrolled in a flagship university using national exam admissions.[56] The effects of the reform on the SES college quality gap are slightly stronger using this intensity of treatment variable. Nonetheless, the similarity of the coefficients in column (B) of Table 5 and column (H) of Table A6 suggests that the benchmark specification captures the first order effects of the exam reform.

Table A7 shows results related to several alternative hypothesis. The table has the same structure as Table 6 but uses different dependent variables. The last row presents estimates of the full-population effect of the reform, and the other rows present separate effects for each SES group.

One potential threat to identification is that the reform affected the probability that students enrolled in any college. For example, some of the negative effects could be due to an increase in the fraction of lower ability students who attended college. The regression in column (A) expands the sample to include all exam takers—not just those who attended college—and the dependent variable is an indicator equal to one if a student attended any college in my records. The estimated effect on any college enrollment is small and insignificant for the full sample, and there are no statistically significant effects on the probability of college enrollment in any SES group. Thus the primary effect of the exam reform was on *where* students went to college, not *whether* they went to college. This may be due to the fact that Colombian college markets have a large, open-enrollment sector where students can enroll if they are not admitted to top colleges. This also parallels the findings in research on other large-scale admission policies that primarily affect admission to selective colleges.

[55] The regressions in columns (E)–(H) of Table A6 replace the region-level fixed effects in specification (11) with municipality-level fixed effects.

[56] Specifically, in columns (G)–(H) Treated_r equals the fraction of college enrollees from the 1998–1999 cohorts in municipality r who enrolled in a flagship that used only national exam scores for admissions. I normalize Treated_r to be mean zero and standard deviation 0.5, which is approximately equal to the standard deviation of Treated_r in the benchmark specification. This makes the coefficients in column (B) of Table 5 and column (H) of Table A6 comparable in magnitude.

For example, other work has found that affirmative action bans (Hinrichs, 2012) and percent plan admission rules (Daugherty et al., 2014) have little effect on the extensive margin of college enrollment.

TABLE A7. Alternative hypotheses

	(A)	(B)	(C)	(D)	(E)	(F)
	Dependent variable					
$\text{Treated}_r \times \text{Post}_t \times \ldots$	Enrolled in any college	Graduating from HS this cohort	Enrolled within one year	Kilometers between HS & college	Enrolled in different region	Employed in formal sector
Top 25 family income	−0.021	−0.013	−0.011	3.712	0.014	−0.019
	(0.014)	(0.016)	(0.007)	(3.709)	(0.010)	(0.016)
Bottom 25 family income	0.001	0.011	0.009	−8.354**	−0.029***	−0.007
	(0.007)	(0.008)	(0.016)	(3.183)	(0.010)	(0.009)
College educated mother	−0.001	−0.005	−0.014	−0.087	0.004	−0.017
	(0.012)	(0.013)	(0.010)	(2.961)	(0.011)	(0.017)
Primary educated mother	0.003	0.010	0.009	−7.382**	−0.028***	−0.004
	(0.008)	(0.009)	(0.013)	(3.370)	(0.009)	(0.007)
High ranked high school	−0.008	−0.018	−0.002	−0.259	0.002	−0.015
	(0.012)	(0.016)	(0.007)	(3.877)	(0.010)	(0.018)
Low ranked high school	0.003	0.013	0.010	−3.845	−0.025**	−0.011
	(0.008)	(0.010)	(0.018)	(4.703)	(0.010)	(0.008)
Male	−0.004	0.006	−0.002	−4.071	−0.018**	−0.006
	(0.011)	(0.008)	(0.010)	(2.975)	(0.008)	(0.010)
Female	−0.003	−0.006	0.002	−3.979	−0.015*	−0.014
	(0.008)	(0.011)	(0.011)	(3.607)	(0.009)	(0.011)
All students	−0.004	−0.000	−0.000	−4.027	−0.017**	−0.010
	(0.009)	(0.009)	(0.010)	(3.218)	(0.008)	(0.010)

Notes: This table reports estimates of θ^y from equation (12) for the dependent variable listed in each column header. The last row shows the estimate of θ^y from the full sample, and the other rows show estimates for separate SES and gender groups.

The sample for the regressions in column (A) includes students in Column (A) of Table 1. The sample for the regressions in columns (B)–(F) includes students in Column (B) of Table 1. Parentheses contain standard errors clustered at the region level.

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

Another possibility is that the reform altered student ability by affecting the probability of retaking the exam or by changing study habits. Columns (B) and (C) show that the reform had little effect on students' likelihood of repeating the exam or delaying college enrollment, two proxies for exam-taking behavior. The dependent variable in column (B) is an indicator equal to one for students who took the exam in their year of high school graduation. The dependent variable in column (C) is a dummy equal to one for students who enrolled in college within one year of taking the admission exam. There are no statistically significant effects on the average or distribution of these variables, which suggests that the reform did not significantly alter the probability that students retook the exam or delayed college entry.

A final possibility is that changes in the distribution of college quality reflect not a resorting of students into colleges, but rather a change in the probability that students stayed in their home region for college. Columns (D) and (E) present evidence that the reform had some effects on student mobility. Column (D) shows an insignificant average effect on the distance between students' high schools and colleges, with a statistically significant average reduction in distance to college for the bottom family income quartile. A similar pattern arises in column (E), which uses a dependent variable equal to one for students who attended college in a region different from that of their high school.

The results in columns (D) and (E) suggest that the reform may have induced some students to attend college closer to home. However, these effects are unlikely to explain the negative graduation and earnings effects. Column (C) of Table 6 shows that there was no statistically significant change in average college quality for the full population, suggesting that students were not systematically picking higher quality colleges in their home region, and that college quotas were not changing dramatically.

Lastly, column (F) shows that no SES group exhibit a significant change in the probability of formal employment 10–11 years after taking the admission exam.

A.8. **Additional robustness tables.**

TABLE A8. Exam reform effects on outcome gaps

	(A)	(B)	(C)	(D)
	Dependent variable: Graduated from college		Dependent variable: Log daily earnings	
	Pre-reform gap	Reform effect	Pre-reform gap	Reform effect
Family income gap	0.146	−0.018*	0.280	−0.005
Top 25 − Bottom 25		(0.010)		(0.009)
Mother's education gap	0.122	−0.017**	0.232	0.001
College − Primary		(0.008)		(0.010)
High school rank gap	0.166	−0.013*	0.269	0.005
High − Low		(0.007)		(0.009)
Gender gap	−0.091	−0.007	0.046	−0.006
Male − Female		(0.007)		(0.007)

Notes: Column (A) depicts the pre-reform graduation gap for each measure of SES and gender, and column (B) shows the estimated effects of the reform. Columns (C)–(D) are analogous using log daily earnings as the dependent variable.

The sample for this table includes students in Column (B) of Table 1. Parentheses contain standard errors clustered at the region level.

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

TABLE A9. Clustered and wild t bootstrap p values

	(A)	(B)
	Dependent variable	
$\text{Treated}_r \times \text{Post}_t \times \ldots$	Graduated from college	Log daily earnings
Top 25 family income	−0.026	−0.022
	(0.005)	(0.023)
	[0.052]	[0.016]
Bottom 25 family income	−0.009	−0.016
	(0.295)	(0.000)
	[0.351]	[0.020]
College educated mother	−0.022	−0.014
	(0.030)	(0.188)
	[0.088]	[0.212]
Primary educated mother	−0.005	−0.015
	(0.511)	(0.002)
	[0.535]	[0.024]
High ranked high school	−0.020	−0.010
	(0.007)	(0.272)
	[0.008]	[0.359]
Low ranked high school	−0.007	−0.015
	(0.487)	(0.008)
	[0.515]	[0.016]
Male	−0.018	−0.018
	(0.009)	(0.002)
	[0.016]	[0.012]
Female	−0.011	−0.012
	(0.210)	(0.073)
	[0.236]	[0.072]
All students	−0.014	−0.016
	(0.058)	(0.004)
	[0.100]	[0.008]

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