Standardized Tests in Holistic University Admissions*

Zachary Bleemer John Friedman †

July 2026

Abstract

We conduct an admissions audit study inside highly-selective private University B to study the role of standardized tests in determining the allocation elite US higher education. Test scores explain 20 percent of systematic admissions variation; within-applicant admission likelihood rises 10 percentage points per 100 SAT points, though the effect is concave for high-scoring students. Math and verbal scores are treated equivalently. Holistic review generates a tradeoff of only 4 average SAT points per 1 percentage point rise in disadvantaged enrollment, far less than single-index admissions models imply. Test-optional admissions does not meaningfully affect enrollment composition by test score or disadvantage.
JEL Codes: I23, I24, J15, L21

*Thanks to Rajeev Darolia and Sahar Parsa for helpful comments; to Luke Mouracade, Thomas Sweeney, and Ariel Zhao for excellent research assistance; and to the admissions staff at University B for support and guidance throughout this project. Any errors that remain are our own.

†Bleemer: Department of Economics, Princeton University. Email: bleemer@princeton.edu. Friedman: Department of Economics, Brown University. E-mail: john_friedman@brown.edu.

1 Introduction

Access to selective universities is a first-order determinant of wages and economic mobility in the United States (Chetty et al., 2020, 2026; Bleemer and Quincy, 2024). While many countries allocate oversubscribed higher education strictly on the basis of pre-college academic preparation, most selective US institutions employ opaque holistic admissions procedures with a broad range of admission criteria (Bleemer, 2023).[1] Students with higher test scores tend to attend more-selective universities (Bleemer and Rothstein, 2025), but there is little evidence on the relative importance or manner of employment of test scores in highly-selective holistic admissions. This study opens the black box of highly-selective admissions to document the role of standardized tests in allocating America's most-selective and highest-value universities.

We conduct an admissions audit study at University B, a high-value private university in America's northeast that rejects 96 percent of its regular applicants. We experimentally manipulate or omit the test scores of 520 applicants and task B's actual admissions 'readers' to evaluate the resulting applications, with each variously-manipulated applicant evaluated by 4-8 different readers. We also manipulate the university's testing policy regime; applications were evaluated on the first two days of our experiment under B's status-quo 'test-optional' policy, in which applicants could choose whether to submit test scores with their applications, and on the third day under a 'test-mandatory' policy in which applicants were required to submit scores.

We first document the importance and use of test scores in highly-selective admissions. Because we observe evaluations of each application at different test scores, we can evaluate their import holding all other application components fixed. Test scores are likely the most impactful component of undergraduate applications: test score variation alone explains about 20 percent of systematic admissions variation at University B, with a relatively higher share (almost one-third) among disadvantaged applicants (which includes racial minorities and low-SES applicants). Within applicant, the likelihood of admission to University B rises by 10 percentage points per 100 SAT points (half of a national standard deviation); additional points on the mathematics and verbal sections of the test have equal effects. While the cross-sectional relationship between test scores

[1]For example, the University of California reports admitting students based on pre-college academic preparation but also special talents in athletics or the arts, strong interest in other cultures, leadership experiences, community service, and other factors.

and admission is convex among high-testing students, however, the causal relationship is concave: University B does not use test scores to distinguish between very high-testing applicants.

Another advantage of our experimental design is that we can identify the average leniency and between-applicant preferences of each of the 21 participating admissions readers, the key arbiters who manifest B's admissions decisions. Treating each reader as a draw from the parameters through which University B's admissions objective is implemented, we identify a new cohort-level parameter that measures the institution's tradeoffs between the average test score of its admitted class and other class characteristics. We find that holistic review sharply dampens the cohort-level relationship between B's average test scores and its disadvantaged share; moving along its objective-inducted cohort-composition gradient, a 10-point increase in B's average SAT score decreases its disadvantaged share by 2.5 percentage points, whereas the relationship would be about four times larger if B employed a simple single-index admission rule.

Finally, we analyze how test-optional admissions reshapes University B access. Withholding a test score only benefits applicants with SAT scores below 1430, the 13th percentile of B's normal second-round applicants. We find weak evidence of a lower cutoff for disadvantaged applicants, perhaps suggesting that readers are less generous in their inference of those applicants' omitted scores. Test-mandatory admissions generates a slightly steeper test-disadvantage gradient in B's admitted cohorts and results in an admitted class predicted to be slightly lower-testing *and* less-disadvantaged.

Because American universities are highly stratified in resources and long-run value-added (Chetty et al., 2020; Bleemer and Rothstein, 2025), there is considerable interest in how access to the highest-resourced (and most-selective) institutions is allocated. Standardized test scores are widely employed in selective university admissions because they are highly predictive of college performance (Douglass, 2007; Lemann, 2000; Bettinger et al., 2013) and long-run labor market outcomes (Friedman et al., 2025). Selection-on-observables designs have shown that test scores are predictive of selective university admission (**?**), though their predictive power declines when universities adopt holistic review **?**. We quantify the overall import of standardized tests in highly-selective admissions and document its non-monotonicity, finding minimal effects among high-scoring applicants (who make up a large share of non-disadvantaged admits) and equal roles for mathematics and verbal exams (despite the stronger predictive power of the former; University of

California, 2020).

There is particular interest in the relationship between standardized tests in admissions and disadvantaged student enrollment because lower-income and underrepresented minority students have much lower average test scores than their more-advantaged peers. [2] A series of conflicting difference-in-difference analyses of annual university enrollment compositions have come to different conclusions regarding the effect of implementing test-optional admissions policies on Black and Hispanic enrollment (Belasco et al., 2015; Saboe and Terrizzi, 2019; Bennett, 2022; **?**); Appendix B conciliates these studies to show that test-optional policies have a small average positive effect on Black enrollment largely driven by less-selective institutions, in line with the very small effect of University B's test-optional policy on its disadvantaged enrollment.[3] Sacerdote et al. (2025) provide evidence that some disadvantaged students see a decline in admissions likelihood after a highly-selective university implemented a test-optional admissions policy because they made poor strategic SAT submission decisions; we XXX.[4] Our study's methodology closely resembles Bastedo et al. (2022), which conducts a stylized audit study at a admissions reader conference estimating the effect of providing readers with contextual information about applicants' high schools and neighborhoods on perceived admissions likelihood; our study goes further by randomly manipulating real applications and matching the environment and conditions of real-world application readings. See also Chen and Jin (2025), Dessein and Kartik (2025), Avery et al. (2025), Bhattacharya and Shvets (2024).

[2] Anti-testing proponents argue that standardized tests bias selective university access away from disadvantaged applicants; e.g. test and retest availability are disproportionately available to higher-income students (Bulman, 2015; Goodman et al., 2020).

[3] Bleemer (2023) and Gandil and Leuven (2022) show that holistic review in general increases disadvantaged enrollment, though to a lesser extent than race-based affirmative action (Bleemer, 2022).

[4] Figure 1 of Sacerdote et al. (2025) shows that the positive cross-sectional relationship between SAT and first-year GPA flattens above 1560, mirroring our finding that the positive relationship between SAT and admission probability flattens around the same point. Dessein and Kartik (2025) predict that the compositional effects of test-optional policies depends on whether test scores provide new information to admissions readers; if score imputation were perfect, there would be no change to admission likelihood (absent salience effects). We instead present evidence admissions officers' beliefs about applicants' scores are biased toward the overall (and perhaps group-specific) mean, thus boosting admission likelihood for low-testing students and decreasing them for high-testing ones.

2 Experimental Design and Data

We conducted an admissions audit study in collaboration with a highly-selective US private university in the United States that we call University B in October 2024. We first randomly selected 10 out of the 18 regions into which B sorts its domestic applications (since application evaluators, called readers, typically evaluate applications from only one region in a given day), weighing regions by their number of 2024 applicants to preserve representativeness. We then randomly sampled 52 regular 2024 applications per region from the 12 percent of applicants who were given 'second reads' by B's admissions office – indicating a non-negligible chance of admission – overweighting both disadvantaged applicants and applicants from disadvantaged high schools.[5]

Assignment to readers featured further layers of randomization. First, we manipulated the standardized test scores of each application. For the SAT score (which has a national mean of about 1000 with standard deviation 200), we implemented four adjustments: a large decrease (roughly -150 to -80 points), a small change (-70 to $+70$), a large increase (roughly $+80$ to $+150$), and a fourth score in the middle of the largest gap between these other three, though this score was omitted in the test-optional regime. The point change was randomly split between the math and reading scores; parallel adjustments were made to ACT scores for students who submitted them. Next, applications were embedded in a rotating packet structure that assigned each application-score to a regional packet. Finally, three packets were as-good-as-randomly assigned to each of 21 readers.[6] Each applications from the 10 regions was assigned to be read 4–5 times under the test-optional regime on the first two days of the experiment – with each reader evaluating 52 applications from a region each day – and applications from 5 of the regions were assigned to be read 4–5 times under the test-mandatory regime on the third day of the experiment.

Table 1 compares the full set of normal applicants – which excludes foreign, early-decision, athlete, and other applicants who experienced non-standard admissions processes) – to the 2,816 second-round applicants who comprised our experimental population and the 519 applicants who were ultimately included in our experiment. Of the $52 \times 3 \times 21$ assigned application reads, 2,717

[5] "Disadvantaged" is defined here as either belonging to an underrepresented minority group or being both first-generation and low-income; disadvantaged high schools have a College Board Landscape score over 50, implying a below-SES-median school.

[6] Readers were assigned regions that they did not evaluate in the actual 2024 admissions cycle, which had occurred six months earlier.

actually occurred; less-experienced or otherwise-slow readers did not complete the remaining reads on the assigned days and were excused from completing them.[7] The weighted sample of selected applicants and the weighted set of reads are both representative on observables of all second-round applicants.

The 21 participating readers comprised most of University B's actual application evaluation staff. They were instructed to evaluate each application as if for 2024 admission to B. Following B's standard admissions process, readers assigned three scores to each application, each on an integer scale of 1–5: an overall rating, an academic rating, and a non-academic rating. We map these ratings into 'predicted admission' ($\hat{A}_{isr}$), the main outcome of interest in our analysis, by first regressing admission on fixed effects of the three ratings and reader fixed effects over the sample of second-round applicants not selected into the experiment and then predicting admission as if assigned the average-leniency reader.

3 Test scores in holistic admissions

3.1 The relevance of test scores in admissions

While far from determinative (as in many other countries; Neilson (2024)), the SAT is likely the most important component of most US universities' holistic admissions process.[8] Table 2 decomposes the variance of predicted admission ($\hat{A}_{isr}$) among the second-round University B applicants in our experimental data using the following model:

$$\hat{A}_{isr} = \alpha_i + \theta_{s_i s} + \delta_r + X_{is} + \epsilon_{isr} \tag{1}$$

where $\hat{A}_{isr}$ depends on applicant i's (experimentally-manipulated) test score s, the variance due to all other components of the application, the variance due to the assigned reader r's leniency, covariances between each of these (and with X_{is}, which contains only an indicator for the test-mandatory regime on day 3), and a residual ϵ_{isr} that largely captures randomness in the admissions

[7] See Figure **??**; the median application was read 5 times.
[8] XXXAdd table with R2's showing large SAT effect on admissions at B.

process.[9] We follow Kline et al. (2020) to estimate unbiased variance components (and standard errors) using a leave-one-out (leave-two-out) estimator.

The SAT variation produced by our experiment – which changed applicants' test scores by an average 61 points (s.d. 46) – explains about 11 percent of variance in predicted admissions, or about a quarter of all admissions variance due to the application materials.[10] All remaining applicant characteristics explain about one-third of total admission variation, with another 5 percent explained by variation in reader leniency. Since the covariance terms are close to 0 by construction – readers are randomly assigned and test scores randomly vary from applicants' true scores – their contribution is minimal.[11] The remaining half of variance in the predicted admission of our experimental application evaluations cannot be explained by any component of the application or by reader leniency.[12]

When predicted admission is broken into its components, we see that the experimental SAT variation explains almost 30 percent of variation in applicants' academic scores but no variation in applicants' non-academic scores. Reader leniency explains none of the variation in academic score but 10 percent of variation in the non-academic score.

The last two columns show that test scores explain almost twice the variance share for disadvantaged applicants as for non-disadvantaged applicants (16 versus 8 percentage points). This likely has two explanations. First, second-round advantaged applicants are more likely to have the potential variation in their SAT score top-coded at 1600, reflecting their lesser possibility of reporting substantially higher test scores. Second, as shown below, disadvantaged applicants are less likely to have test scores near the 1500-1550 range where the causal relationship between test scores and University B admissions is particularly strong.

[9] Applications with experimentally-missing SAT scores are assigned separate fixed effects by the applicant's true SAT score s_i to reflect readers differently interpreting missing SAT scores for different applicants; see Figure 3(a) below.

[10] The SAT variance component reflects the effect of our experimental test score variation, which was considerable (± 150 points). This is likely an upper bound on plausible applicant manipulability of test scores. Hoxby (2025) similarly finds that SAT scores explain about a quarter of explainable variation in UNC Chapel Hill admissions decisions using a cross-sectional Shapley decomposition.

[11] The applicant-score covariance explains about 10 percent of variation in $Acad_{is}$ because the SAT scores are centered at applicants' true scores, but explains little other rating variation; see Table A-3.

[12] The potential explainable variance of $\hat{A}_{isr}$ is bounded above due to its discreteness; $\hat{A}_{isr}$ is predicted from three five-value ratings and takes 73 unique values in the data. We simulate a continuous $\hat{A}_{isr}$ by predicting it using covariates and then predict a rounded version of the prediction by the continuous measure. We find that discreteness can only explain 2.5 percentage points of the residual variation in Table 2.

3.2 The use of test scores in admissions

The relationship between test scores and college admissions is strongly positive. The admission rates of students within 20 points of a 1300 (1.5 standard deviations above mean), 1400, 1500, and 1600 were 1.5, 2.7, 5.1, and 17.0, respectively. Figure 1(a) shows that the SAT-admission relationship is concave, with a stronger relationship among higher-SAT students, and persists among second-round applicants (though less strongly): higher-testing applicants are both more likely to make it into University B's second admissions round and are more likely to be admitted among second-round applicants.

Because our experiment varies applicants' SAT score across application readers, we can separately identify the between-applicant cross-sectional relationship from the *within-applicant* causal effect of reporting a different SAT score on college admission. We estimate:

$$\hat{A}_{isr} = \beta SAT_s + \zeta_i + \eta_r + X_{is} + \epsilon_{is} \tag{2}$$

where ζ_i absorbs applicant fixed effects – holding fixed everything in the application but the SAT score – η_r absorbs average reader leniency, and X_{is} includes cluster-by-day fixed effects to absorb spurious variation from cross-application spillovers and testing regime.

When we restrict to the one 'true' observation for each application (with approximately their action SAT score) and estimate Equation 2 without individual fixed effects α_i, the resulting $\beta^{between} = 0.079(0.024)$: cross-sectionally, applicants with 100 higher SAT scores are 7.9 percentage points more likely to be admitted to B. This cross-sectional relationship is the sum of the causal effect of having a higher SAT score and 'selection bias' resulting from higher-SAT students also having other characteristics that increase or decrease their admission likelihood at B. Figure 1(a) shows that estimating Equation 2 over the experimental sample results in $\beta^{within} = 0.104(0.008)$: increasing one's SAT score by 100 points increases admissions likelihood by over 10 percentage points. This implies that the 'selection bias' is positive: net of SAT score and among second-round University B applicants, the other characteristics of high-SAT applicants are less valued by B's admissions procedure than those of low-SAT applicants.[13]

The relationship between SAT and admissions is non-monotonic. Figure 1(b) shows estimates

[13] Figure XXX shows applicant FE against SAT score.

by within-sample SAT tercile and binscatter estimates for six SAT bins.[14] Higher SAT scores are most impactful for students in the middle tercile, between 1500 and 1550; students' causal admission likelihood rises 10 percentage points in that span. The relationship is strongly concave, however, in contrast with the concavity of the between-student relationship; increasing one's SAT score from 1550 to 1600 increases predicted admission by only one percentage point. The high-SAT convexity of the observational relationship between SAT score and admissions is wholly explained by 'selection bias,' other non-SAT characteristics of higher-testing students that are positively correlated with both SAT score and admission to B.

There is little other measurable heterogeneity in the test-admission relationship. Observable student characteristics do not meaningfully mediate the causal relationship between SAT score and admission. Figure 1(c) shows that the causal effect of having a higher SAT score on University B admissions is approximately equal for disadvantaged (10 percentage points per hundred SAT points) and non-disadvantaged (11 percentage points) applicants; Figure XXX shows further similarity between other groups.[15] There is also little difference between the causal effect of higher SAT verbal scores relative to higher math scores (holding the other fixed); despite the latter being more predictive of later-life outcomes **?**, if anything the effect of higher verbal scores (12 percentage points) is higher than that of math scores (9 percentage points), though the two are statistically indistinguishable.[16]

3.3 How test scores trade off against other applicant characteristics

How do universities employ SAT scores relative to other applicant characteristics when constructing their admission classes? Prior studies of universities' preferences over applicant characteristics have typically summarized them by $\hat{\lambda} = \frac{\hat{\beta}^D}{\hat{\beta}^S}$ (Espenshade et al., 2004), where:

$$A_i = \beta^S SAT_i + \beta^D Disadv_i + \phi' X_i + \epsilon_i \tag{3}$$

[14] The lowest-SAT binscatter bin mostly comprises applications with experimental SAT scores lower than their true scores; the highest-SAT bin mostly comprises applicants with experimentally-increased SAT scores. Figure A-1 shows that there is little heterogeneity in the causal effect of raising versus lowering applications' SAT scores, suggesting these bins remain directly comparable.

[15] We obtain these coefficients by estimating Equation 2 over the respective restricted sample.

[16] We estimate Equation 2 where SAT_s is a vector of both SAT scores, estimating the effect of each score holding the other fixed.

The resulting $\hat{\lambda}$ is conventionally interpreted as the SAT-equivalent advantage associated with $Disadv_i$.[17] It is therefore a partial-equilibrium estimate for individual applicants. At University B, $\hat{\lambda} \approx 100$ for disadvantaged students (see Table A-1).[18]

However, holistic admissions weighs many correlated characteristics without employing a single admissions index and likely values the resulting class composition directly. As a result, institutions' preferred admission cohorts need not lie on a common SAT–disadvantage margin. We therefore estimate a cohort-level parameter: across realizations r of University B's admissions objective that produce higher- or lower-testing cohorts, how does the cohort's composition change?

Our experimental setting permits identification of a four-characteristic preference vector C_r^R for each application reader, including their overall leniency as well as the relative weights that they place on applicants' academic ($\hat{\gamma}_{is}^A$), disadvantage ($\hat{\gamma}_i^D$), and non-academic ($\hat{\gamma}_i^N$) strength using an extension of Equation 1:[19]

$$\hat{A}_{isr} = \alpha_i + \theta_{s_i s} + \delta_r + \delta_r^A \hat{\gamma}_{is}^A + \delta_r^D \hat{\gamma}_i^D + \delta_r^N \hat{\gamma}_i^N + X_{is} + \epsilon_{isr} \tag{4}$$

where $C_r^R = \{\hat{\delta}_r, \hat{\delta_r}^A, \hat{\delta}_r^D, \hat{\delta}_r^N\}$ is defined by the Bayesian-shrunk estimates of each $\hat{\delta}_r$ vector.[20] Admissions readers are trained to administer the admissions rule of their institution, and we treat each as representing a draw from the parameters through which University B's objective is implemented. We apply each reader's four-parameter admission rule to the full second-round applicant

[17] Following public discussion and academic discourse at the time (e.g. **???**), we focus on B's tradeoff between high-testing students and disadvantaged students (who generally have lower test scores).

[18] Prior estimates of $\hat{\lambda}$ have generally come from institutions implementing race-based affirmative action; e.g. estimates of $\hat{\lambda}$ for Black students at highly-selective universities were 300–400 in the 1980s and 1990s (Kane, 1998; Espenshade and Radford, 2009). Due to the 2023 *SFFA v. Harvard* Supreme Court decision, University B did not provide race-based preferences in 2024. Zitzewitz (2025) finds that $\hat{\lambda}$ rises from 57 to 120 for Black applicants to UCLA when that school implemented a new holistic review policy (under an affirmative action ban).

[19] We construct indices of each applicant's academic, disadvantage, and non-academic strength as:

$$\hat{A}_{isr} = \beta^A X_{is}^A + \beta^D X_i^D + \nu_{s_i - s} + \epsilon_{isr}$$

$$\hat{\gamma}_{is}^A = Stand(\hat{\beta}^A X_{is}^A) \qquad \hat{\gamma}_i^D = Stand(\hat{\beta}^D X_i^D) \qquad \hat{\gamma}_i^N = Stand(\hat{A}_{is} - \hat{\beta}^A X_{is}^A - \hat{\beta}^D X_i^D)$$

where $\nu_{s_i - s}$ includes fixed effects for the difference between i's true and experimental SAT score; X_{is}^A includes applicants' academic characteristics (SAT math and verbal scores and high school GPA); X_i^D includes applicants' disadvantage characteristics (Black, Hispanic, low parental income, first-generation, high school landscape value, financial aid flag, and high school share XXX); and $Stand(\cdot)$ standardizes the vector. The student's non-academic strength is the application's strength residualized on covariates.

[20] The fifth column of Table 2 shows that these reader characteristics explain over 10 percent of the non-random variation in $\hat{A}_{isr}$, with about two-thirds coming from leniency and one-third coming from relative preferences.

pool and calculate the resulting cohort's mean SAT (S_r) and disadvantaged share (D_r).[21] We then define the objective-induced cohort-composition gradient as

$$\kappa = \frac{\text{Cov}(S_r, D_r)}{\text{Var}(S_r)} \tag{5}$$

so that κ summarizes how B's disadvantaged share co-moves with mean SAT as the university's full bundle of priorities varies.

Figure 2 shows that $\hat{\kappa} = -0.25$: for each percentage point increase in B's disadvantaged share, its average SAT score will fall by $\frac{1}{0.25} = 4$ points.[22] If University B admitted applicants using a single index, on the other hand – assigning disadvantaged applicants a fixed value of $\hat{\lambda} \approx 100$ SAT points – then the marginal disadvantaged admit would have λ fewer SAT points than the marginal non-disadvantaged admit, and variation in the admissions rule would imply $\hat{\kappa} = \frac{-100}{\hat{\lambda}} \approx -1$.[23] The fourfold discrepancy between these estimates shows that the applicant-level bonus λ is not a sufficient statistic for the cohorts the university builds: University B's preference for high-testing students is far less at odds with its preference for disadvantaged enrollment than the single-index benchmark predicts. This does not imply that University B places little value on disadvantage; rather, its sizable applicant-level preference for disadvantage does not translate into a comparably strong cohort-level SAT–disadvantage relationship. While λ describes how disadvantage affects individual applicants' evaluation, κ describes how the institution's admission objectives are manifested in equilibrium cohort composition.

Figure 2 further contextualizes this revealed SAT-disadvantage admission tradeoff in several ways. First, University B's cohort-composition gradient lies well within the SAT-disadvantage feasibility frontier among second-round applicants, reflecting B's pursuing many objectives beyond

[21] In particular, for each reader $\rho \in R$, we measure for each second-round applicant:

$$E[A_{isr}|r=\rho] = \hat{A}_{isr} + (\hat{\delta}_\rho - \hat{\delta}_r) + \sum_{Z \in \{A,D,N\}} (\hat{\delta}^Z_\rho \hat{\gamma}^Z_{is} - \hat{\delta}^Z_r \hat{\gamma}^Z_{is})$$

restricting to applications evaluated (in the first round) by the 15 readers who also participated in our experiment. We then estimate, e.g., $S_r = \frac{1}{N}\sum_i E[A_{isr}|r=\rho]SAT_i$.

[22] Hoxby (2025) constructs another κ-like equilibrium statistic: the relationship between average SAT score and disadvantaged share between alternative 'feasible' single-index admissions policies designed to increase disadvantaged enrollment using race-neutral policies (though she does not involve university objectives in her analysis). She estimates a URM-SAT slope of about -0.6, likely close to $\frac{-100}{\hat{\lambda}}$ in that setting (though this is not reported).

[23] See Appendix C for a discussion of the standard applicant-independent benchmark admissions model and an extension that matches our observed effects.

these two dimensions. Second, the composition of actually-admitted second-round students – which also reflects post-read discussion and negotiation between admissions staff – has a slightly higher average SAT score and share disadvantaged than would be expected given applicants' $\hat{A}_{isr}$ (1530 with 23 percent disadvantaged), which is consistent with the post-read admission process placing slightly higher relative weight on those characteristics. Third, the overall composition of University B's admitted class – which includes students admitted through its early decision process and other special admitted students – falls on the gradient but with a higher disadvantaged share (and lower average SAT score) than any reader provided by their preferences, further validating the gradient estimation and suggesting that other constraints (like the applicant pool or narrowly-tailored admissions slots for specific applicants) may require these other admissions processes to relatively favor disadvantaged applicants.

While we focus primarily on the SAT-disadvantage tradeoff, our gradient approach can be generalized to map the institution's broader compositional signature. Figure XXX shows that κ is near 0 for legacy admissions and XXX.

4 Holistic review with test-optional admissions

The share of US universities requiring standardized tests in admissions rose from below 10 percent in the 1930s to near-universality in the 1980s, but steadily declined after 2000 before sharply declining in the Covid-19 pandemic; as of 2025, it remains close to 10 percent (**?**). How do admissions readers change their evaluation of SAT scores – and evaluate applications without scores – under a test-optional admissions policy?

Our experiment includes two days of test-optional application evaluation and one day of test-mandatory evaluation. We estimate:

$$\hat{A}_{isr} = \alpha_i + \delta_r + \omega^S s + \omega^M Missing_{is} + \omega^{SM} s \times Missing_{is} + \omega^X X_{is} + \epsilon_{isr} \tag{6}$$

where s is the experimental SAT score and is replaced by the applicant's true SAT score if the experimental SAT is omitted ($Missing_{is} = 1$) and X_{is} includes just a day-3 fixed effect. When we estimate Equation 6 including an interaction term between s and an indicator for being evaluated

under the test-optional regime, the resulting coefficient of 0.1 (s.e. 1.5) rejects any meaningful difference in SAT score's causal effect with the test-mandatory regime $\omega^S = 10.5$; we omit this term below.

Figure 3(a) visualizes the estimated within-applicant effect of omitting one's SAT score on the applicant's predicted admission to University B by the applicant's true SAT score.[24] Unsurprisingly, admissions readers penalize high-testing applicants for withholding their scores (likely believing their scores to be lower) and reward low-testing applicants for withholding. The x-intercept of 1434 (s.e. 30) implies that 1430 – the 13th percentile of second-round applicants' SAT scores (and the 37th percentile among all normal applicants) – is the average cutoff point; readers admit below-cutoff applicants as if they have higher-than-true scores, so applicants with scores below this cutoff benefit from withholding their scores.

The figure also shows similar estimates separately for disadvantaged and non-disadvantaged applicants. They have near-identical slopes but a large (though statistically-noisy) gap in x-intercepts: there is weak evidence ($t = 1.2$) that application readers are relatively less generous to SAT-omitting disadvantaged applicant, such that any such applicant with a test score above 1390 benefits from submitting the score (relative to 1460 for advantaged applicants).[25]

We also estimate how test-optional policies shift universities' overall revealed admissions objective by comparing University B's cohort-composition gradient between SAT score and disadvantage under each admissions regime. We estimate Equation 4 interacting $\theta_{s_i s}$ and the components of C_r^R – the relative admissions advantages assigned to different SAT scores and readers' preferences – with the test policy regime. Figure 3(b) shows that the test-mandatory regime has a 10-percent steeper gradient than the test-optional regime: test-mandatory cohorts with higher average test scores tend to have slightly lower disadvantaged shares than under test-optional admissions, with a steeper gradient by about 10 percent. As a result, the expected average test score and disadvantage share of test-optional cohorts are both slightly higher than under test-mandatory cohorts, though the differences are very small (e.g. 0.15 percentage points of share disadvantage). This is in line with our synthesis of prior research on the net compositional effects of test-optional policies presented in Appendix **??**.

[24] The line has y-intercept $\hat{\omega}^M$ and slope $\hat{\omega}^{SM}$.

[25] XXXSacerdote

5 Conclusion

References

Avery, C., L. Shi, and P. Magouirk (2025). Test-optional college admissions: Act and sat scores, applications, and enrollment changes. NBER Working Paper 34260.

Bastedo, M., K. Glasener, K. C. Deane, and N. Bowman (2022). Contextualizing the sat: Experimental evidence on college admission recommendations for low-ses applicants. Educational Policy 36(2), 282–311.

Belasco, A., K. Rosinger, and J. Hearn (2015). The test-optional movement at america's selective liberal arts colleges: A boon for equity or something else? Educational Evaluation and Policy Analysis 37.

Bennett, C. T. (2022). Untested admissions: Examining changes in application behaviors and student demographics under test-optional policies. American Educational Research Journal 59(1), 180–216.

Bettinger, E., B. Evans, D. O. Staiger, and D. Pope (2013). Improving college performance and retention the easy way: Unpacking the act exam. American Economic Journal: Economic Policy 5(2), 26–52.

Bhattacharya, D. and J. Shvets (2024). Inferring trade-offs in university admissions: Evidence from cambridge. Journal of Political Economy 132(11).

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

Bleemer, Z. (2023). Affirmative action and its race-neutral alternatives. Journal of Public Economics 220, 104839.

Bleemer, Z. and S. Quincy (2024). Changes in the college mobility pipeline since 1900.

Bleemer, Z. and J. Rothstein (2025). The meritocratic consensus and stratification in higher education. Manuscript.

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

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. Journal of econometrics 225(2), 200–230.

Chen, M. and G. Jin (2025). Voluntary report of standardized test scores: An experimental study. NBER Working Paper 33660.

Chetty, R., D. Deming, and J. Friedman (2026). Diversifying society's leaders? the determinants and causal effects of admission to highly selective private colleges. The Quarterly Journal of Economics 141, 51–145.

Chetty, R., J. Friedman, E. Saez, N. Turner, and D. Yagan (2020). Income segregation and intergenerational mobility across colleges in the united states. Quarterly Journal of Economics 135(3), 1567–1633.

Dessein, W. and N. Kartik (2025). Test-optional admissions. American Economic Review 115(9), 3130–3170.

Douglass, J. A. (2007). The California idea and American higher education: 1850 to the 1960 master plan. Alhoda UK.

Espenshade, T. J., C. Y. Chung, and J. L. Walling (2004). Admission preferences for minority students, athletes, and legacies at elite universities. Social Science Quarterly 85(5), 1422–1446.

Espenshade, T. J. and A. W. Radford (2009). No longer separate, not yet equal: Race and class in elite college admission and campus life. Princeton, NJ: Princeton University Press.

Felegi, B. (2026). The impacts of removing college entrance exams: evidence from the test-optional movement. Southern Economic Journal 92(4), 935–955.

Friedman, J. N., B. Sacerdote, D. O. Staiger, and M. Tine (2025). Standardized test scores and academic performance at ivy-plus colleges. AEA Papers and Proceedings 115, 676–681.

Gandil, M. and E. Leuven (2022). College admission as a screening and sorting device. IZA Discussion Paper 15557.

Goodman, J., O. Gurantz, and J. Smith (2020). Take two! sat retaking and college enrollment gaps. American Economic Journal: Economic Policy 12(2), 115–158.

Hoxby, C. (2025). Affirmative action: Logic, methods, and evidence. Economics of Education NBER Summer Institute 15557.

Kane, T. J. (1998). Racial and ethnic preferences in college admissions. Ohio St. LJ 59, 971.

Kline, P., R. Saggio, and M. Solvsten (2020). Leave-out estimation of variance components. Econometrica 88(5), 1859–1898.

Lemann, N. (2000). The big test: The secret history of the American meritocracy. Macmillan.

Neilson, C. (2024). The rise of coordinated choice and assignment systems in education markets around the world. World Bank Development Report: Background Papers 2024.

Saboe, M. and S. Terrizzi (2019). Sat optional policies: Do they influence graduate quality, selectivity or diversity? Economics Letters 174, 13–17.

Sacerdote, B., D. O. Staiger, and M. Tine (2025). How test optional policies in college admissions disproportionately harm high achieving applicants from disadvantaged backgrounds. National Bureau of Economic Research Working Paper 33389.

Zitzewitz, E. (2025). The academic consequences of affirmative action bans combined with diversity targets. NBER Working Paper 33563.

Note: Cite Conlon & Conlon, and the new test-optional theory paper

Table 1: Descriptive Statistics

	Applicant Mean Characteristics				Experimental Statistics		
	All Normal Applicants	Second Round	RCT Sample Raw	RCT Sample Weighted	Mean	St. Dev.	IRR
Female (%)	62.2	50.7	53.4	53.5	53.6		
Disadvantaged (%)	27.3	20.5	25.0	18.6	19.6		
SAT Score	1472	1537	1534	1540	1540	47	
Math	738	774	772	775	775	33	
Verbal	735	763	763	765	766	26	
Omitted SAT (%)	38.2	16.2	22.4	17.5	18.3		
HS GPA	3.70	3.76	3.73	3.73*	3.74	0.23	
Overall Rating[1] (1–5)	1.71	3.08	3.11	3.10	2.61	0.56	0.33
Academic Rating[1] (1–5)	3.52	4.67	4.63	4.70	4.33	0.52	0.46
Non-Acad. Rating[1] (1–5)	3.59	4.33	4.34	4.36	4.08	0.64	0.39
Pred. Admission[1] ($\hat{A}_{isr}$; %)	9.6	37.0	37.9	38.2	27.8	18.0	0.34
Admitted (%)	3.8	31.3	28.5	28.8	26.2		
Matriculated (%)	1.4	11.2	10.8	10.8	9.7		
Observations	23,606	2,816	519	519		2,717	

Note: The first four columns present the average characteristics of normal 2024 applicants to University B (which excludes early-decision, foreign, athlete, and Questbridge applicants; applicants who applied to small specialized programs or who were flagged by university administration; and applicants missing key demographic information), the applicants who are selected into the second round of the admissions procedure (our experimental population), the applicants who were randomized into our experiment, and the weighted characteristics of those applicants (taking into account our oversampling of disadvantaged applicants and applicants from disadvantaged high schools). Weighted experimental means that differ from those of all second-round applicants at 5 percent significance are marked with an asterisk (*). The last three columns present the mean and standard deviation of our experimental sample – where each observation is an application reader's evaluation of one of the applicants in the RCT sample – along with the intra-reader reliability (IRR) for readers' evaluations of applications. 'Disadvantaged' is defined as being Black, Hispanic, or both first-generation and an Expected Family Contribution (EFC) below $15,000 (which approximately corresponds to a parental income of $110,000); predicted admission ($\hat{A}_{isr}$) is the predicted likelihood of admission given the applicant's overall, academic, and non-academic ratings assuming they were evaluated by the average-leniency reader, as estimated using the second-round evaluations of second-round applicants excluded from the experiment. High school GPA is unweighted (out of 4). IRR is defined as the unweighted absolute-agreement intraclass correlation (ICC) from a crossed (applicant $\times$ reader) random-effects model, net of the experimental SAT (or missing SAT) and experiment day. [1]Applicant ratings and predicted admission reflect first-round evaluations in the first four columns and experimental evaluations in the latter columns.

Source: University B Administrative Data.

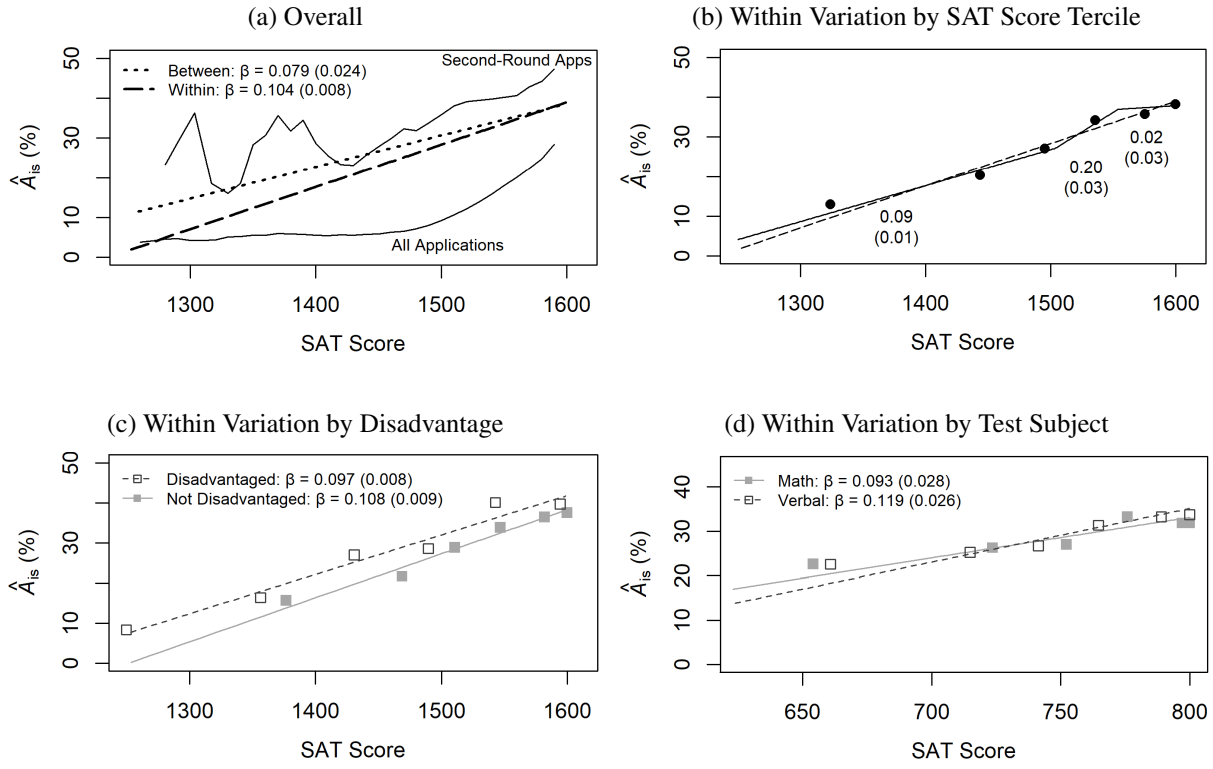
Table 2: Decomposition of Variance in Highly-Selective Holistic Admissions

Sample:	All Students					Disad.	Adv.
	$\hat{A}_{isr}$	$Rate_{is}$	$Acad_{is}$	$NonAcad_{is}$	$\hat{A}_{isr}$	$\hat{A}_{isr}$	$\hat{A}_{isr}$
Percent of Variance Explained:							
Applicant	31.0	31.2	21.3	40.3	30.7	30.3	31.7
	(1.5)	(3.2)	(5.9)	(7.8)	(3.6)	(3.4)	(1.6)
Test Score	10.6	9.0	28.4	-0.1	6.6	15.5	8.4
	(1.4)	(1.5)	(2.5)	(0.5)	(2.7)	(3.8)	(1.4)
Reader	4.5	6.4	1.7	9.8	4.4	5.7	5.3
	(0.8)	(0.9)	(0.4)	(1.4)	(0.8)	(1.4)	(1.0)
Reader Prefs.					2.0		
					(0.5)		
Covariances	0.0	-0.8	12.4	-0.6	4.4	-5.5	2.1
	(1.2)	(1.3)	(1.1)	(0.6)	(5.2)		
Residual	53.3	53.8	36.3	50.5	51.3	52.7	52.2
	(1.2)	(2.9)	(5.8)	(7.6)	(0.9)	(2.2)	(1.2)
# of Students			516			130	386
Observations			2,713			711	1999

Note: This table presents decompositions of the variance of predicted admission ($\hat{A}_{isr}$) and the overall, academic, and non-academic rating of experimental 2024 University B applicant evaluations into shares explained by the full application portfolio (excluding the SAT score), the experimentally-manipulated SAT score, the assigned application reader, covariances between these terms (which includes the main effect and covariances with whether the evaluation occurred on the third test-mandatory day of the experiment), and the residual. Variance components are estimated using the leave-out estimator of Kline et al. (2020), which replaces each observation's squared-error contribution with an unbiased leave-one-out estimate; four observations with leverage 1 are excluded. The fifth column further includes terms capturing the assigned reader's relative preferences for disadvantaged students, strong academic students and applicants with strong non-academic credentials; the sixth and seventh columns restrict the decomposition to disadvantaged and non-disadvantaged applicants, respectively. 'Disadvantaged' is defined as being Black, Hispanic, or both first-generation and an Expected Family Contribution (EFC) below $15,000 (which approximately corresponds to a parental income of $110,000); predicted admission ($\hat{A}_{isr}$) is the predicted likelihood of admission given the applicant's overall, academic, and non-academic ratings assuming they were evaluated by the average-leniency reader, as estimated using the second-round evaluations of second-round applicants excluded from the experiment.

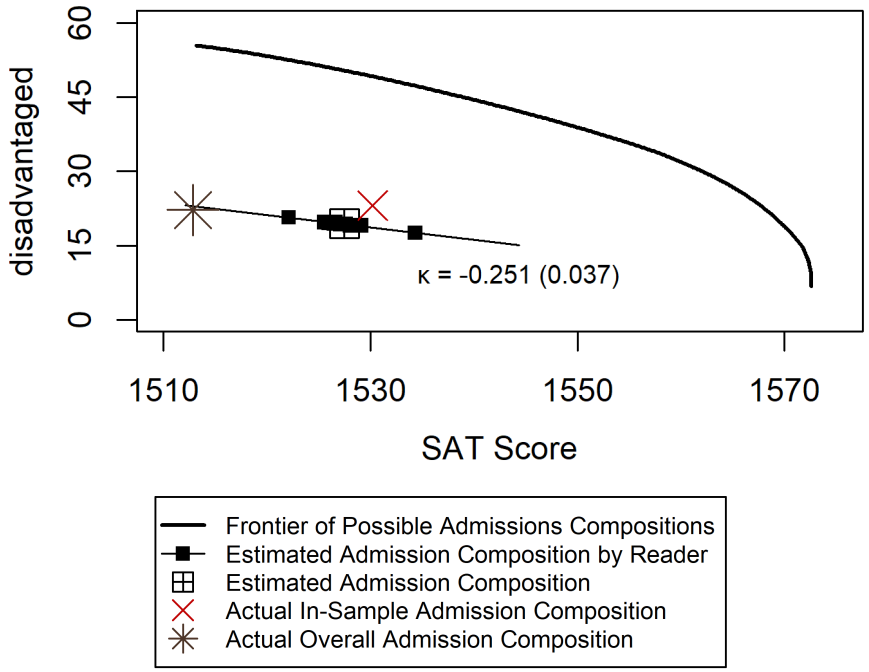
Source: University B Administrative Data.

Figure 1: The Causal Effect of Test Scores on Highly-Selective University Admissions



Note: **Panel (a)** presents the average predicted admission likelihood ($\hat{A}_{isr}$) of all 2024 applicants to University B and among applicants who are selected into the second round of the admissions procedure (our experimental population) by actual SAT score; the between-applicant relationship between SAT score and $\hat{A}_{isr}$ in the experimental sample (restricting to the one experimental evaluation of the 'true' application per applicant); and the within-applicant relationship between SAT score and $\hat{A}_{isr}$ in the experimental sample, estimated with applicant fixed effects and scaled to the mean applicant. The legend includes coefficients (and standard errors in parentheses) of the slopes, which are not statistically distinguishable. **Panel (b)** splits the within-applicant slope by SAT tercile and shows non-parametric estimates of the relationship by SAT sextile; the three slopes are all statistically distinguishable at the 1 percent level. **Panel (c)** splits the within-applicant slope by whether the student is disadvantaged and shows non-parametric estimates of each relationship by within-group SAT sextile. The legend includes coefficients (and standard errors in parentheses) of the slopes, which are not statistically distinguishable. **Panel (d)** splits the within-applicant slope into math and verbal SAT components and shows non-parametric estimates of each relationship by component sextile. The legend includes coefficients (and standard errors in parentheses) of the slopes, which are not statistically distinguishable. 'Disadvantaged' is defined as being Black, Hispanic, or both first-generation and an Expected Family Contribution (EFC) below $15,000 (which approximately corresponds to a parental income of $110,000); predicted admission ($\hat{A}_{isr}$) is the predicted likelihood of admission given the applicant's overall, academic, and non-academic ratings assuming they were evaluated by the average-leniency reader, as estimated using the second-round evaluations of second-round applicants excluded from the experiment. Standard errors are clustered by region-grouping (40 total clusters). Source: University B Administrative Data.

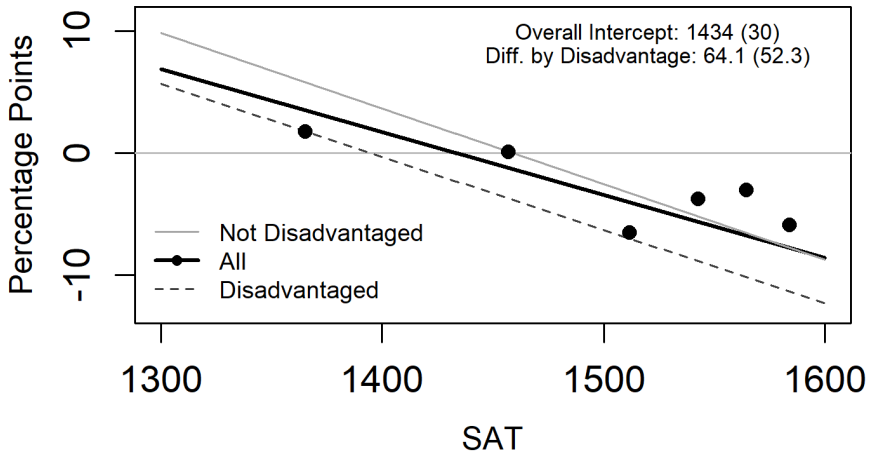
Figure 2: Cross-Reader Tradeoff Between SAT Score and Disadvantage



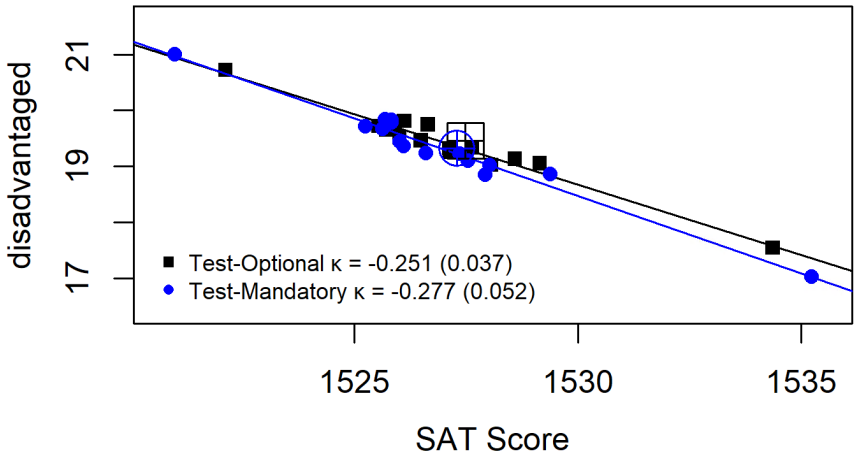
Note: This figure presents the predicted mean SAT score and share disadvantaged of admitted University B applicants if each of the 21 admissions readers were to conduct admissions alone under the status-quo test-optional regime; the best-fit line between these reader-specific admission compositions; the overall predicted in-sample admissions composition using the experimental evaluations; the actual composition of all admitted second-round applicants; and the actual composition of all admitted students at University B, including students admitted outside of regular second-round admissions (like early-decision students and athletes); and the frontier of possible SAT-disadvantaged combinations, measured by admitting the N top-scoring second-round applicants (where N is the number of actually-admitted second-round applicants) and then sequentially rejecting the lowest-scoring non-disadvantaged applicant in favor of the highest-scoring disadvantaged applicant. Predicted admission composition is the characteristic's weighted average by the applicant's predicted admission; predicted admission ($\hat{A}_{isr}$) is the predicted likelihood of admission given the applicant's overall, academic, and non-academic ratings assuming they were evaluated by the average-leniency reader, as estimated using the second-round evaluations of second-round applicants excluded from the experiment. Predicted admission is adjusted to being as if each reader conducted all admissions by subtracting off the estimated reader characteristics of the assigned reader – overall leniency and relative preference for disadvantage, academics, and non-academics – and adding in those of each reader in turn. 'Disadvantaged' is defined as being Black, Hispanic, or both first-generation and an Expected Family Contribution (EFC) below $15,000 (which approximately corresponds to a parental income of $110,000). Source: University B Administrative Data.

Figure 3: The Effect of Test-Optional Admissions

(a) Student Decision



(b) Institutional Decision



Note: **Panel (a)** presents the estimated change in admissions likelihood if a student with each SAT score were to withhold their score from consideration under an SAT-optional admissions policy, overall and by whether the student is disadvantaged, assuming linearity by SAT score or (for the overall estimates) allowing the estimate to vary by SAT sextile. The presented coefficients (and Delta-method standard errors in parentheses) estimate the overall x-intercept and the difference between the x-intercepts of the estimates by disadvantage. See equation 6. **Panel (b)** replicates the SAT-disadvantaged tradeoff presented in Figure 2 (see the note to that figure for details) and adds a further set of predicted mean SAT score and share disadvantaged compositions of admitted University B applicants if each of the 21 admissions readers were to conduct admissions alone *and admissions were conducted under a test-mandatory regime*, holding second-round applications fixed (with SAT-omitting students submitting their scores) and allowing reader characteristics (overall leniency and relative preference for disadvantage, academics, and non-academics) and admissions likelihoods by SAT score to adjust in the new regime (see equation 4). The best-fit line between these SAT-mandatory reader-specific admission compositions is also shown; the slope parameters are shown in the legend. 'Disadvantaged' is defined as being Black, Hispanic, or both first-generation and an Expected Family Contribution (EFC) below $15,000 (which approximately corresponds to a parental income of $110,000). Source: University B Administrative Data.

Online Appendix

Standardized Tests in Holistic Review

Zachary Bleemer and John Friedman

July 2026

Table of Contents

Appendix A: Experimental Design 2
Appendix B: Effect of Pre-Covid Test-Optional Policies on Student Composition 6
Appendix C: Testing Models of Holistic University Admissions 13
Other Appendix Figures and Tables 17

Appendix A: Experimental Design

We conducted an admissions audit study in collaboration with University B, a highly-selective private US university. University B has two admissions cycles, a smaller "early decision" cycle in November in which students are permitted to apply to only that institution and commit to enrolling if admitted, and a "regular admissions" cycle in January. In the regular admission cycle, each application is assigned to a first 'reader' and evaluated, and then about 12 percent of more-promising applications are assigned second readers and are evaluated in a second round. Applications are block-assigned to readers by geography and are evaluated in both rounds by readers assigning three numbers to each between 1 and 5: an overall rating, an academic rating, and a non-academic rating. Highly-rated applications are then discussed in meetings between admissions officers, with committee votes ultimately determining admission to B.

University B offered that 20 professional admissions readers could evaluate undergraduate admissions files for three days as part of an admissions audit study, with a typical reader able to evaluate about 50 applicants per day. The experiment was conducted over three days in late October 2024, just before evaluations for the 2025 early decision cycle began. University B had implemented a test-optional admissions policy in 2024 – applicants could choose whether to submit standardized test scores as part of their application – but it had already announced that it would implement a test-mandatory policy in 2025. We told the admissions readers to evaluate applications as they had in 2024 (that is, under the test-optional policy) on the first two days of the experiment, and then told them on the third day to evaluate applications as they would under B's new test-mandatory policy.

We selected and constructed the applications that each reader evaluated as follows. We began with the admissions files of all actual 2024 applicants to B. We then imposed the following data restrictions:

Restriction	Omitted	Resulting Sample Size
All Applicants		48,896
Regular Admissions Cycle	6,239	42,657
Not Foreign	9,621	33,036
Not a Recruited Athlete	29	33,007
Not Questbridge	4,229	28,778
Not Part of Special Program	3,330	25,448
No Missing Variables	1,842	23,606
Second Round	20,790	2,816

We omitted early-decision and foreign applicants to focus on the admissions process for regular domestic applicants. We omitted recruited athletes because they are admitted through a different admissions process, as are applicants through the Questbridge program (which facilitates selective university admissions for participating lower-income students) and applicants to special programs at Brown, which include the dual degree program with the Rhode Island School of Design, Brown's

Program in Liberal Medical Education (an eight-year joint undergraduate and medical degree), and applicants flagged by Brown administration for special consideration. We omit students whose admission file does not include the College Board Landscape score used to measure the disadvantaged status of their high school, their race, or their parental income or education.

Finally, we omit students who were rejected prior to being evaluated in the second round. This eliminates applications with essentially no chance of being admitted to University B, considerably increasing the observed variation in applicants' assigned ratings. As a result, our audit study most closely mirrors the second-round regular admissions process at University B.

Next, we sample applications to be used in our experiment. First we randomly select 10 of the 18 regions into which B sorts its domestic applications, weighting each super-geo by its number of 2024 applicants.[26] The resulting regions are numbered G1 to G10.

Next we sampled 52 applications per region. University B did not permit direct heterogeneity analysis by applicant race, but permitted analysis by two alternative measures: applicants from high-adversity high schools (those with above-median Landscape scores) and 'disadvantaged' applicants, which is defined as being minority-race (Black, Hispanic, or Native American) or being both first-generation (that is, neither parent graduated from college) and having a federal Expected Family Contribution (EFC) below $15,000.[27] In order to increase our statistical power to test for differences between more- and less-disadvantaged students, we assign sample weight 2.5 to applicants from high-adversity high schools and offsetting lower sample weight to applicants who neither attended such a high school nor was 'disadvantaged'.

For each region G#, we next randomized the 52 applicants selected from that region into four 13-applicant clusters, which we call G#-a to G#-d. Each cluster in turn was replicated into four 13-applicant sub-clusters, in which we manipulated their standardized test scores as follows:[28]

- G#-x-i: We replace the SAT score with a lower score. For applicants with an SAT score above 1500, their score declines by $60-100$ points. For applicants with an SAT score below 1300, their score declines to a score between 1200 and their true score. For applicants in between, their score declines by $80-150$ points.

- G#-x-ii: We replace the SAT score with a similar score. For applicants above 1500, their score changes by $-30-30$ points. For applicants below 1300, their score changes to somewhere between 1300 and 1400. For applicants in between, their score changes by $-70-70$ points.

- G#-x-iii: We replace the SAT score with a higher score. For applicants above 1500, their

[26] We exclude New York City because the Regents Exam Scores are printed on students' transcripts and may be incongruous with experimentally-assigned SAT scores, and Florida because their transcripts are difficult to read for new readers.

[27] The average family income of a college student with an EFC between $13,000 and $17,000 was $111,000, among respondents to the 2019 NPSAS.

[28] We conducted a small pilot study among senior University B admissions officers using manipulated New York City applications and found that it was important to not manipulate high-scoring applicants' scores too far downward, because the resulting applications seemed impossible to admissions readers. The rules below reflect this feedback.

score increases by $50 - 90$ points. For applicants below 1300, their score increases to 1400 plus $0 - 100$ points. For applicants in between, their score increases by $80 - 150$ points.

- G#-x-iv: On days 1 and 2 of the experiment (when the readers are operating under a test-optional policy), these applicants omit their scores. On day 3, we assign these applicants the median score between the other two scores above assigned to the applicant that are furthest from each other.

All values are rounded to the nearest 10 points and censored at 1600. SAT scores are reported overall and separately for math and verbal; we randomly distribute the adjustment between the two, drawing uniformly from $[0.25, 0.75]$ and rounding to the nearest 10 points.[29] About one-quarter of applicants submit ACT scores instead of SAT scores; we adjust these scores by the intervals $[-6, -3]$, $[-2, 2]$, and $[3, 6]$, applying the same score changes overall and to each separate component and censoring the overall and component scores at 36.

We then recombined these pieces to define four packets for each region, rotating the applications selected so that one-quarter each have a large increase, small change, large decrease, or omitted score:

- G#-1 = G#-a-i + G#-b-ii + G#-c-iii + G#-d-iv
- G#-2 = G#-a-ii + G#-b-iii + G#-c-iv + G#-d-i
- G#-3 = G#-a-iii + G#-b-iv + G#-c-i + G#-d-ii
- G#-4 = G#-a-iv + G#-b-i + G#-c-ii + G#-d-iii

We then assign these packets to readers in a manner that results in a well-connected lattice of reads of each set of packets:

	Test-Optional		Test-Mand.		Test-Optional		Test-Mand.
	Day 1	Day 2	Day 3		Day 1	Day 2	Day 3
Reader 1	G1-1	G6-2	G3-3	Reader 11	G2-2	G1-3	G5-4
Reader 2	G2-1	G7-2	G4-3	Reader 12	G3-2	G4-3	G2-4
Reader 3	G3-1	G8-2	G5-3	Reader 13	G6-1	G8-4	G1-2
Reader 4	G4-1	G9-2	G1-3	Reader 14	G7-1	G5-3	G3-2
Reader 5	G5-1	G10-2	G2-3	Reader 15	G10-1	G9-4	G4-2
Reader 6	G6-3	G5-4	G2-1	Reader 16	G1-2	G3-3	G4-4
Reader 7	G7-3	G4-4	G1-1	Reader 17	G4-2	G6-4	G3-4
Reader 8	G8-3	G3-4	G5-1	Reader 18	G5-2	G2-3	G1-4
Reader 9	G9-3	G2-4	G4-1	Reader 19	G8-1	G10-4	G2-2
Reader 10	G10-3	G1-4	G3-1	Reader 20	G9-1	G7-4	G5-2

[29] Neither subscore can be over 800; any subscores over 800 are censored, with the remaining points added to the other subscore.

Readers are were assigned numbers randomly, but were switched if they were experimentally assigned an application that they had actually read in the 2024 admissions cycle. Notice that all ten regions were evaluated on the first two days but only the (randomly-chosen) first five were evaluated on the third day, in order that each evaluated application on the third day was evaluated four times.

When the experiment was actually administered, 21 readers took part. The 21st reader joined only for days 1 and 3 and read arbitrarily-chosen packets of applications that we had not assigned to be read on that day. Moreover, several readers ended up reading different packets than those they were supposed to be assigned. The following table shows, for each reader, which reader's packet they read on each day of the experiment, denoting in bold the cases where readers ended up reading packets that they had not been intentionally assigned:

Reader	Day 1	Day 2	Day 3	Reader	Day 1	Day 2	Day 3
1	1	1	1	12	**15**	12	12
2	2	2	2	13	**20**	13	13
3	3	3	3	14	14	14	14
4	4	4	4	15	**12**	15	15
5	5	5	5	16	16	16	**19**
6	**7**	**8**	6	17	**X**	**6**	17
7	**6**	7	7	18	**18**	18	18
8	8	**17**	8	19	**18**	19	**16**
9	9	9	**11**	20	**13**	20	20
10	10	10	**9**	21	G8-4	**X**	G6-3
11	11	11	**10**				

Note: Bold entries show where the assignment does not match the readers' implementation. For example, on Day 1, Reader 7 read the assignment of Reader 6, and vice versa. **X** denotes a day when the reader completed no reads. We gave the 21st reader no assignments, so their self-assignments are displayed instead.

Moreover, many readers failed to evaluate their full set of applications on each day. The admissions office told them to skip any applications that were not completed on their assigned day. Some readers also flagged (and did not evaluate) a small number of applications in which the true SAT score was reported on the student's transcript or elsewhere in the application (and was thus incongruous with the reported score) or where the score implausible given other applicant characteristics. The readers completed the following number of evaluations:

Reader	Day 1	Day 2	Day 3	Reader	Day 1	Day 2	Day 3
1	47	51	52	12	38	45	45
2	52	52	52	13	39	45	48
3	50	50	45	14	49	49	51
4	43	52	52	15	45	46	50
5	47	52	52	16	46	26	31
6	52	52	52	17	0	32	24
7	45	45	49	18	47	52	17
8	35	42	32	19	27	52	45
9	52	52	51	20	44	44	44
10	45	45	45	21	31	0	8
11	52	50	52				

Appendix B: Effect of Pre-Covid Test-Optional Policies on Student Composition

In 1998, the enrollment-weighted share of universities that required undergraduate applicants to submit standardized test scores peaked at 90 percent (and 92 percent among private universities; Bleemer and Rothstein, 2025). It then steadily declined to 77 percent in 2019 before precipitously declining during the Covid-19 pandemic, reaching 8 percent in 2024. Before 2020, most schools that ended their test-mandatory admissions policies implemented test-optional policies instead. What is the average effect of implementing a test-optional admissions policy on universities' enrollment composition – particularly among disadvantaged students, who tend to have lower average test scores – and how does that effect differ by institutional type?

Four published studies have used the same dataset – Integrated Postsecondary Education Data System (IPEDS), an annual federal survey of US colleges and universities – and the same research design (staggered difference-in-differences) to estimate the effect of implementing a test-optional admissions policy on student composition. Belasco et al. (2015) and Saboe and Terrizzi (2019) found statistically-noisy negative point estimates on the share of the incoming freshman cohort who are Black or Hispanic, the best available measure of disadvantage available in IPEDS.[30] Bennett (2022) reports that test-optional policies slightly *increase* Black and Hispanic enrollment, by 0.7 percentage points ($t \approx 2$); Felegi (2026) reports a larger 1.4 percentage point increase ($t \approx 3$).

Table BB-1 summarizes the differences in specification between these four studies. Belasco et al. (2015) restrict their sample to 180 selective liberal arts colleges – which includes most of the early institutions which adopted test-optional policies – and included a series of institution-level control variables in their two-way fixed effect specifications. Saboe and Terrizzi (2019)

[30]IPEDS has only reported the annual share of first-year students at each institution who receive Pell grants, a proxy for low-income students, since 2008. Because most students stay four or more years at an institution, overall student composition statistics are rolling averages of incoming cohorts over many years and are difficult to study in an impulse-response fashion using staggered difference-in-differences.

Table BB-1: Comparison With Published Event Study Estimates

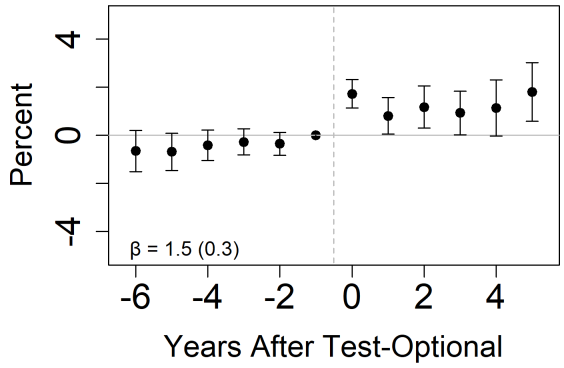
	Belasco *et al*	Saboe & Terrizzi	Bennett	Felegi	Bleemer & Friedman
Study Year	2015	2019	2022	2026	2026
Institutions	180 selective LA colleges	All four-years	Private four-years	All four-years	Private four-years
Year Range	1992-2010	2009-2014	2001-2015	2001-2018	2001-2020
Control Schools	All	All	2016-2020 Treated Inst.	All	All
Covariates	Enrollment, Tuition, Exp. Per Student, and No-Loan Indic.	Exp. and Endowment Per Student	Enrollment, Tuition, Exp. Per Student, and No-Loan Indic.	Enrollment, Tuition, 6-Year Grad. Rate	None
Weights	None	None	MDM Propensity	Callaway and Sant'Anna	None
Fixed Effects	Institution + Year	Institution + Year	Institution + Year + Inst. Time Trends	Institution + Year	Institution + Year
Coef. for:					
Log # of Applications	0.089 (0.043)	0.012 (0.034)	0.031 (0.018)	—	—
Share Enroll. URM (%)	-0.4 (0.5)	-1.6 (2.0)	0.7 (0.4)	1.4 (0.5)	1.5 (0.3)
Share Enroll. Female (%)	—	—	6.0 (1.2)	—	0.9 (0.3)

Note: Characteristics of prior and current studies' estimation of staggered difference-in-difference models of the effect of implementing test-optional admissions policies on university characteristics, all estimated using IPEDS data, and the resulting estimated effects on the log number of applications and the percent of first-time full-time freshmen who are Black, Hispanic, or Native American (URM) or female. Source: Authors' observations and replication.

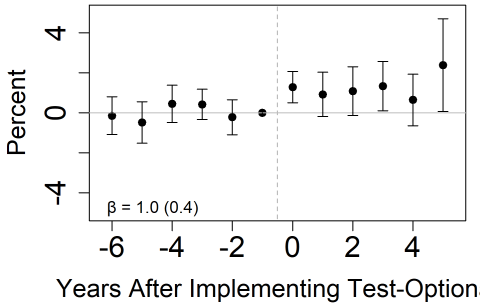
expand their sample to all institutions but restrict to a much smaller number of years. Bennett (2022) was a substantial step forward in this literature, particularly in his improving the set of institution-year pairs that implemented test-optional policies relative to the earlier studies. He restricted to private four-year institutions – which included all events in his period – restricted the control schools to late-adopters, and estimated two-way fixed effect models with institution-level controls, propensity weights, and institution-specific time trends. Finally, Felegi (2026) returned to the simpler specification of Belasco et al. (2015) using the identified events from Bennett (2022) and estimated a newer difference-in-difference model (Callaway and Sant'Anna, 2021) in place of two-way fixed effects.

Figure BB-1: Average Effect of Test-Optional Admissions on URM Share of Enrollment

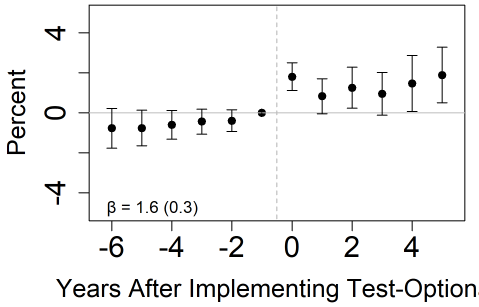
(a) All Schools



(b) Most/Highly Competitive Schools



(c) Less-Competitive Schools



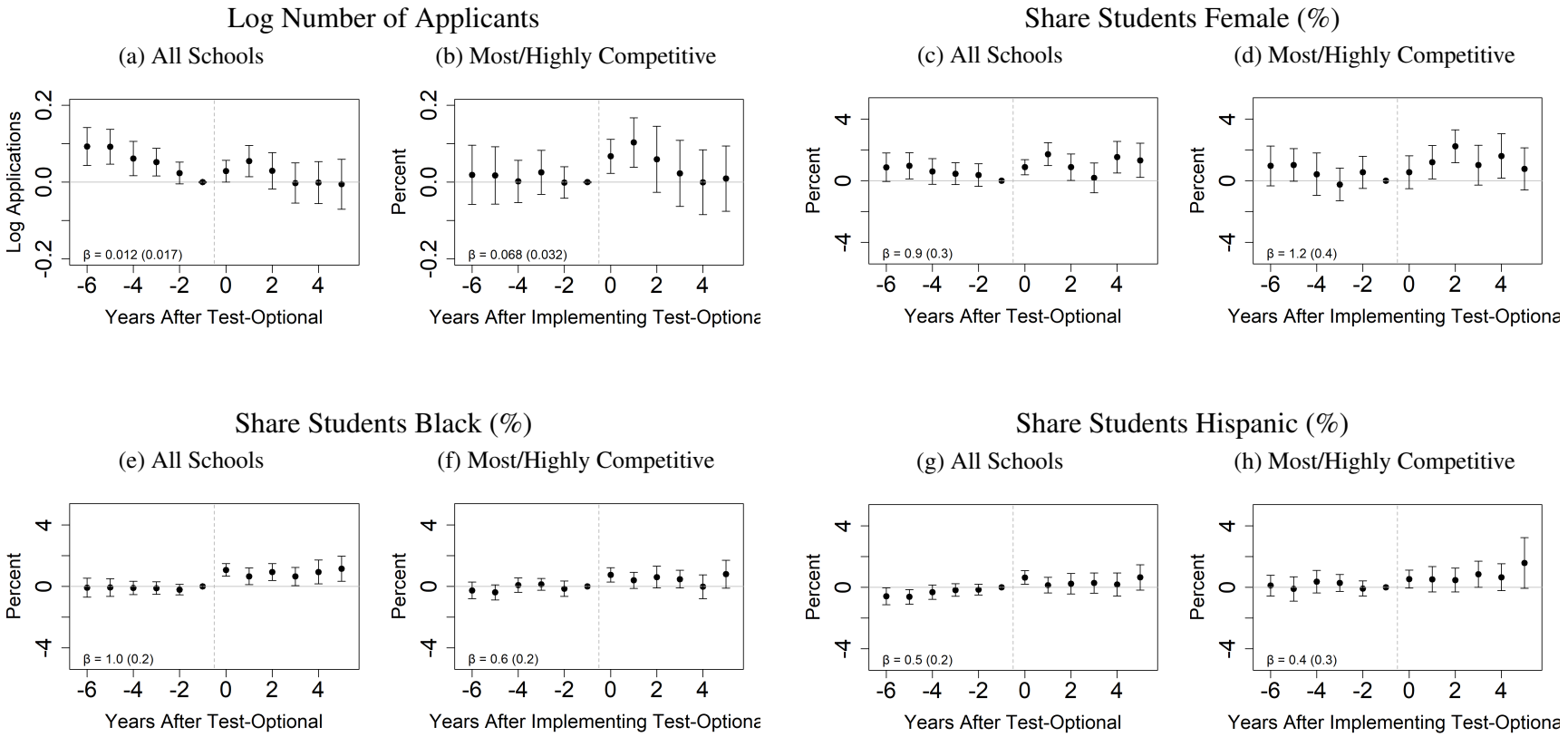
Note: Estimates from an institution-year fixed effect model of the percent of first-time full-time freshmen who are Black, Hispanic, or Native American (URM) on the number of years before or after the institution implemented a test-optional admissions policy, estimated over all 2001–2020 private institutions in the US and splitting the sample into institutions in the first two Barron's selectivity categories ('Most Competitive' or 'Highly Competitive') in 2004 and all other institutions. The coefficient in the bottom-left corner comes from a version of the model where the three years prior to policy implementation ($t \in [-3, -1]$) are omitted and the three subsequent years ($[t \in 0, 2]$) are combined into a single indicator; the coefficient on that indicator is reported (with standard error in parentheses). Standard errors are clustered by institution. Treatment is defined following Bennett (2022). Source: IPEDS.

We arbitrate between these various models and findings by adopting what we believe to be the most parsimonious and transparent combination of these various approaches. As show in Table BB-1, we follow Bennett (2022) in his event definitions and full sample (restricting to private universities).[31] However, we estimate two-way fixed effect models with no further controls, comparing the student compositions of institutions that adopt test-optional admissions policies to other private institutions as simply as possible.[32]

[31]Thanks to Bennett for generously providing us with his dataset of test-optional events.

[32]We also include events between 2016 and 2020 in our analysis. Our baseline estimates are unbalanced: some schools open and close in the sample period, and we restrict the data to 2001-2020 to avoid the 2021 Covid-19 disruption (when nearly all schools implemented test-optional policies). There are 227 total events in our sample, with the number of events per year increasing over time; 35 are at most- or highly-competitive schools. The total sample

Figure BB-2: Effect of Test-Optional Policies on Selective and Non-selective Schools



Note: Estimates from an institution-year fixed effect model of the log number of applications and the percent of first-time full-time freshmen who are female, Black, or Hispanic on the number of years before or after the institution implemented a test-optional admissions policy, estimated over all 2001–2020 private institutions in the US and restricting the sample to institutions in the first two Barron's selectivity categories ('Most Competitive' or 'Highly Competitive') in 2004. The coefficient in the bottom-left corner comes from a version of the model where the three years prior to policy implementation ($t \in [-3, -1]$) are omitted and the three subsequent years ($[t \in 0, 2]$) are combined into a single indicator; the coefficient on that indicator is reported (with standard error in parentheses). Standard errors are clustered by institution. Treatment is defined following Bennett (2022). Source: IPEDS.

The resulting baseline event study estimates are shown in Figure BB-1(a). There is slight evidence that institutions that implement test-optional admissions are already seeing a slight increase in URM enrollment in the years prior to policy implementation, possibly because test-optional admissions policies are implemented by institutions simultaneously implementing other policies designed to diversify their student bodies. Nevertheless, there is a clear sharp increase in URM enrollment in the year that test-optional admissions is implemented. Comparing the three years before and after policy implementation, URM enrollment tends to increase by about 1.5 percentage points (s.e. 0.3).

When we split the sample into institutions deemed 'most- or highly-competitive' by Barron's in 2004 (35 of which implement test-optional policies in our sample period) and less-competitive institutions, we see that while the estimated treatment effect is larger at less-competitive schools, the two are not statistically distinguishable. The 1 percentage point increase in URM enrollment confirms the direction of our estimate of the effect of test-optional admissions at University B (see Figure 3) but provides a larger magnitude. This may suggest that our focus on second-round normal admissions evaluations leads to an underestimate of the effect of test-optional admissions on URM enrollment at University B; perhaps differences in first-round admissions (which determines who is evaluated in the second round) or in other parts of the admissions process (like athletic admissions) also contribute to changes in URM enrollment.

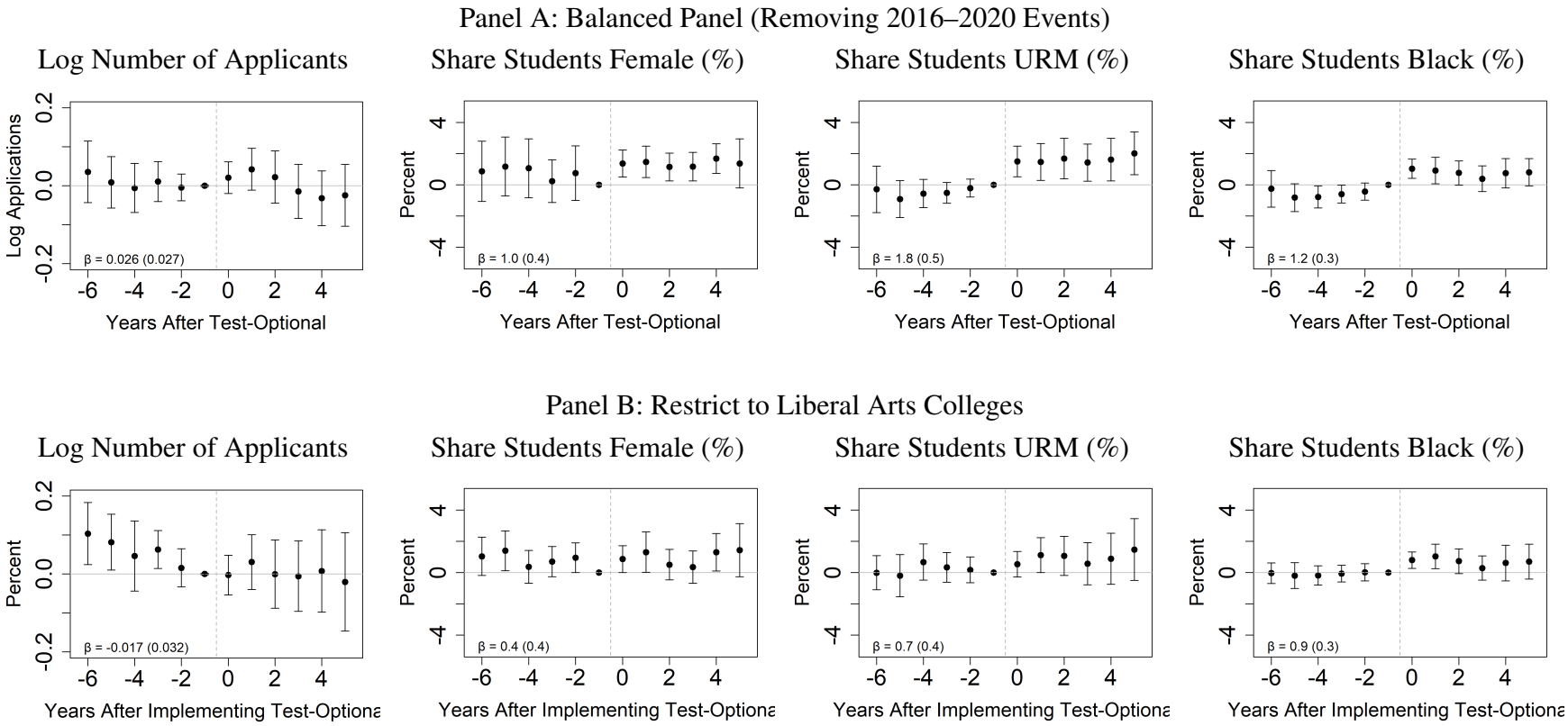
Figure BB-2 shows our baseline event study estimates for other outcomes. Institutions that implement test-optional policies tend to have relatively falling numbers of applications, though this is wholly driven by less-competitive institutions; more-competitive institutions see a 5–10 percent increase in applications following test-optional implementation. Test-optional schools see a small increase in female enrollment. About two-thirds of the increase in URM enrollment comes from increased Black enrollment, which exhibits highly parallel pre-trends both overall and among more-selective institutions.

Figures BB-3 and BB-4 investigate the sources of differences between our estimates and those presented in prior studies. Restricting to a balanced panel – which requires dropping any university that does not appear the IPEDS data in every year between 2001 and 2020 and excluding the institutions that implemented test-optional admissions before 2007 or after 2015, resulting in 97 remaining events – slightly increases the estimated effect of test-optional implementation on URM enrollment, though this may be due to a prior trend (Figure BB-3, Panel A). The balanced panel, then, does not explain our larger estimate than that from Bennett (2022). Nor does Bennett (2022)'s use of covariates: Figure BB-4 shows that none of the covariates meaningfully changes in the years before or after test-optional implementation, suggesting that they play a minor role in his analysis. Instead, the inclusion of institution-specific time trends and the restriction to a small set of control institutions likely explains the small difference.

Panel B of Figure BB-3 shows that the estimated effect of test-optional admissions on URM enrollment is only about half the average size when restricted to liberal arts colleges, though it remains positive. The noisy negative estimate from Belasco et al. (2015) is likely the result of

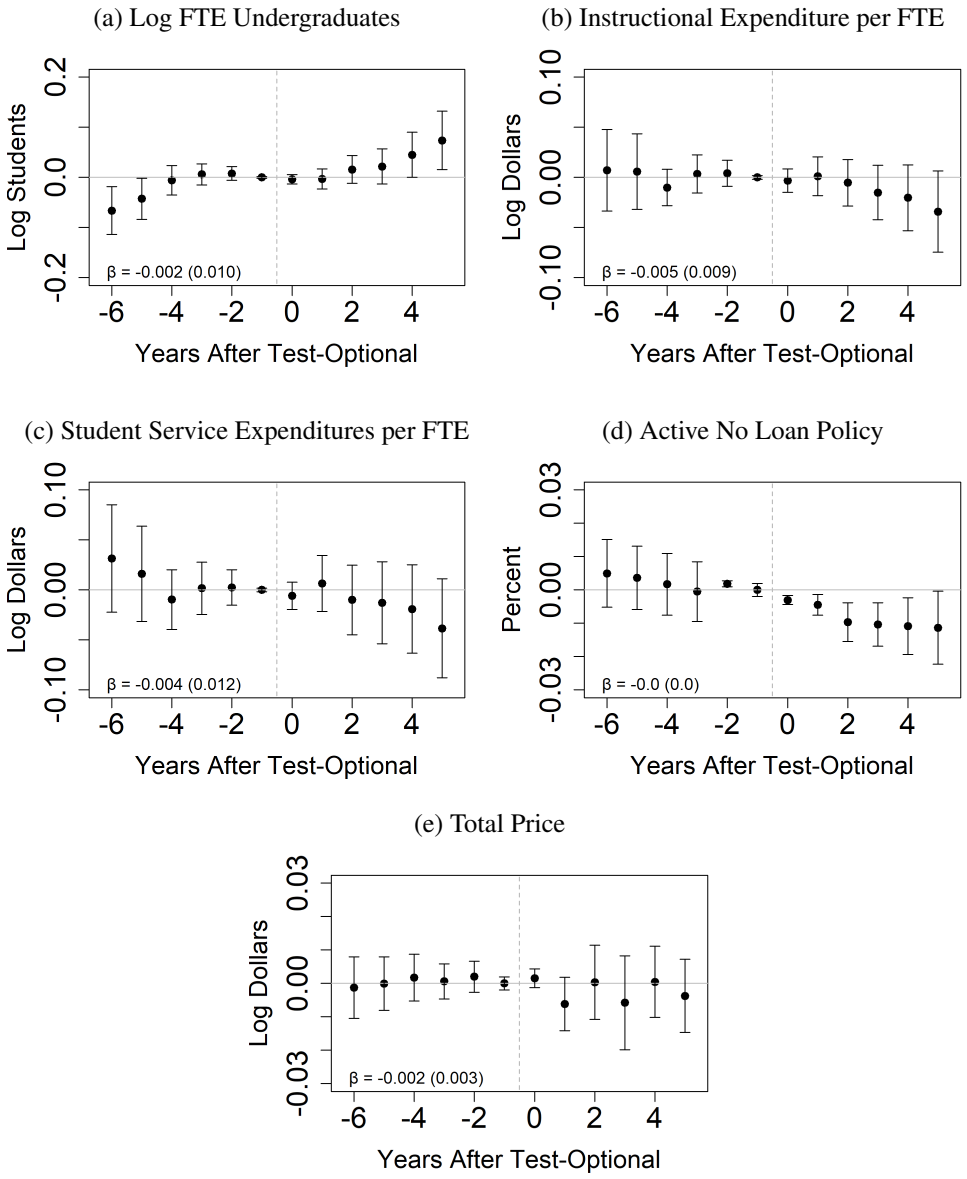
includes 20,431 observations of 1,218 institutions between 2001 and 2020.

Figure BB-3: Effect of Test-Optional Policies on Selective and Non-selective Schools



Note: Estimates from an institution-year fixed effect model of the log number of applicants or the percent of first-time full-time freshmen who are female, Black, or Black, Hispanic, or Native American (URM) on the number of years before or after the institution implemented a test-optional admissions policy, estimated over **Panel A:** all 2001–2020 private institutions in the US that appear in IPEDS in every year between 2001 and 2020, and omitting any institution that implemented a test-optional admissions policy before 2007 or after 2015; or **Panel B:** restricting to institutions with 2000 Carnegie classification 31 and in the top three categories of 2004 Barron's competitiveness (following Belasco et al., 2015). The coefficient in the bottom-left corner comes from a version of the model where the three years prior to policy implementation ($t \in [-3, -1]$) are omitted and the three subsequent years ($[t \in 0, 2]$) are combined into a single indicator; the coefficient on that indicator is reported (with standard error in parentheses). Standard errors are clustered by institution. Treatment is defined following Bennett (2022). Source: IPEDS.

Figure BB-4: Balance Figures of Bennett's Covariates



Note: Estimates from an institution-year fixed effect model of the number of full-time-equivalent (FTE) students, per-FTE instructional expenditure, per-FTE student services expenditure, an indicator for whether the institution has a stated 'no-loan' policy, and the log total sticker price of attendance on the number of years before or after the institution implemented a test-optional admissions policy, estimated over all 2001–2020 private institutions in the US and splitting the sample into institutions in the first two Barron's selectivity categories in 2004 and all other institutions. The coefficient in the bottom-left corner comes from a version of the model where the three years prior to policy implementation ($t \in [-3, -1]$) are omitted and the three subsequent years ($[t \in 0, 2]$) are combined into a single indicator; the coefficient on that indicator is reported (with standard error in parentheses). Standard errors are clustered by institution. Treatment is defined following Bennett (2022). Source: IPEDS.

smaller effects from liberal arts college implementations in the 1990s and from small improvements in event year measurement identified by Bennett (2022).

Appendix C: Testing Models of Holistic University Admissions

[NOTE: This text is almost straight from Claude; it needs to be worked on, and a version may eventually go into the main text]

This appendix presents a simple model of admissions that clarifies the relationship between the applicant-level bonus λ and the cohort-composition gradient κ. The model shows that under a benchmark in which the university values each applicant independently of the rest of the admitted cohort, κ is bounded by two channel-specific slopes; we then test those bounds and discuss which departures from the benchmark the estimates do and do not support.

An applicant-independent benchmark

A university admits A of N second-round applicants with characteristics (s_i, d_i, u_i): an SAT score, a disadvantage indicator, and a vector of other application characteristics. In the benchmark, the university's valuation of each applicant is independent of the composition of the rest of the admitted cohort: reader r admits the A applicants with the highest index

$$V_{ir} = s_i + \lambda_r \, d_i + w_r' u_i, \tag{CC-1}$$

normalizing the SAT weight to one, so that λ_r is the SAT-point bonus for disadvantage and w_r the (SAT-denominated) weights on other characteristics. Following the text, each reader's weight vector (λ_r, w_r) represents a draw from the distribution of parameters implementing University B's objective. Note that this benchmark is fully 'holistic' in the sense of valuing the entire application; its restrictive feature is that each applicant's value does not depend on who else is admitted. Let S_r and D_r denote the mean SAT and disadvantaged share (in percentage points) of reader r's cohort, let κ be the cross-reader regression slope of D_r on S_r, and let Cov_m and Var_m denote moments among applicants near the admission margin.

Cross-reader variation in the admissions rule moves cohorts through two distinct channels.

Proposition 1. *In the applicant-independent model:*
(i) *Dispersion across readers in the disadvantage bonus λ_r moves cohorts along a locus of slope $-100/\lambda$. A marginal increase in the bonus preserves capacity through one-for-one swaps: a disadvantaged applicant whose SAT score is λ points below the marginal non-disadvantaged admit replaces that admit, raising D by $100/A$ and lowering S by λ/A.*
(ii) *Dispersion in any other weight moves cohorts along a locus of slope equal to the margin pool gradient*

$$b_m \;\equiv\; 100 \cdot \frac{\mathrm{Cov}_m(d, s)}{\mathrm{Var}_m(s)}, \tag{CC-2}$$

provided that other characteristics are uncorrelated with disadvantage conditional on SAT at the margin: a perturbation to the weight on u_j tilts admission toward high-u_j applicants, moving S in proportion to $\mathrm{Cov}_m(s, u_j)$ and dragging D along at the rate at which disadvantage covaries with SAT among marginal applicants.

(iii) *The cross-reader gradient κ is a variance-weighted average of these channel slopes and therefore lies between $-100/\lambda$ and b_m, with weights given by each channel's share of the cross-reader variance in cohort mean SAT.*

The conditional-independence requirement in (ii) is testable: Table A-1 shows that $\hat{\lambda}$ is essentially unchanged by the inclusion of a rich covariate set, indicating that the bonus is specific to disadvantage rather than mediated by correlated characteristics. The benchmark thus delivers a sharp prediction: κ must lie in the interval bounded by the bonus channel's slope and the pool's margin geometry, and its location within that interval reveals where the university's internal variation lives.

Empirical results

We estimate all three quantities, localizing the bonus and pool moments to the admission margin by weighting with $\hat{A}(1-\hat{A})$, the derivative of admission probability with respect to the admission index.[33]

- **Bonus channel.** The margin-weighted bonus is $\hat{\lambda}_m = 137$ (Table A-1, column XXX), implying a bonus-channel slope of $-100/\hat{\lambda}_m \approx -0.73$.
- **Pool channel.** The margin-weighted regression of disadvantage on SAT among second-round applicants yields $\hat{b}_m = -0.27$ (Table A-1, final columns), essentially unchanged by controls. The margin-weighted disadvantaged SAT gap (roughly 70 points) is close to the unconditional second-round gap, so the pool's geometry at the margin is not composition-neutral: marginal disadvantaged applicants genuinely score lower.
- **Observed gradient.** $\hat{\kappa} = -0.25$ (0.04), clustered by region-grouping and stable across policy regimes (-0.25 under test-optional and -0.28 under test-mandatory evaluation; Figure **??**b).

The observed gradient sits at the pool-channel boundary of the benchmark's predicted interval $[-0.73, -0.27]$. In a cluster bootstrap of the full estimation pipeline, $\hat{\kappa}$ is flatter than $-100/\hat{\lambda}_m$ in 99 percent of replications: we reject the hypothesis that cohort composition trades off against SAT at the rate implied by the applicant-level bonus. By contrast, $\hat{\kappa}$ is statistically indistinguishable from $\hat{b}_m$ (flatter in 75 percent of replications, a two-sided p of roughly 0.5): cohort co-movement is fully consistent with the pool's margin geometry.

Proposition 1(iii) says this configuration arises when the cross-reader variance in the admissions rule is concentrated in the non-bonus dimensions. The reader characteristics estimated in Section 3.3 confirm this directly. Figure **??** shows that readers differ substantially in leniency

[33] Applicants at the admission margin are those whose admission responds to perturbations of the rule; under a logistic mapping from index to admission probability, the response weight is exactly $\hat{A}(1-\hat{A})$. Results are similar using indicator windows of intermediate admission probability.

($\hat{\sigma} = 5.2$, SE 1.4) and in their relative preferences for academic ($\hat{\sigma} = 2.3$, SE 0.8) and non-academic ($\hat{\sigma} = 2.2$, SE 0.8) strength, but display no detectable dispersion in their treatment of disadvantage ($\hat{\sigma} = 0.8$, SE 1.2, indistinguishable from zero). The dimension of University B's objective with the least internal variation is precisely the price of disadvantage. Because the variation that generates cohort differences is nearly orthogonal to disadvantage, cohorts slide along the pool's margin geometry at b_m, and the bonus channel's much steeper slope never expresses in cohort composition.

Together these estimates form an overidentified and mutually consistent account: a large, uniform applicant-level bonus ($\lambda_m \approx 137$, with negligible cross-reader dispersion); substantial cross-reader variation in other priorities; and a cohort-composition gradient pinned to the pool's margin gradient rather than to the bonus. The failure of $-100/\lambda$ as a prediction for κ is thus fully explained within the applicant-independent benchmark: the extrapolation errs not because the bonus is misestimated, but because it assumes the bonus is what varies, when the university's internal variation is in its other priorities.

Departures from applicant-independence

Several natural models relax the assumption that applicants are valued independently of the cohort; our estimates discipline them.

Composition preferences. Suppose the university values the disadvantaged share directly, with diminishing marginal value: utility includes a term $A \cdot G(D)$ with $G' > 0 > G''$. Admission still follows a cutoff rule, but with an endogenous effective bonus $\lambda_r^{\text{eff}} = \gamma_r + G'(D_r)$ that rises whenever the cohort's disadvantaged share slips. Such preferences act as a restoring force on D_r, damping the expression of all channels in the disadvantaged share and pushing κ toward the flat side of the benchmark interval. Our estimates cannot reject this extension — the point estimate $\hat{\kappa} = -0.25$ sits at (in fact marginally beyond) the flat boundary of the benchmark's prediction, which is the direction composition-anchoring would push — but they do not require it. Notably, a composition preference operating on the institution's readers rather than its cohorts offers a natural interpretation of the pattern above: the one dimension in which University B's readers exhibit no detectable disagreement is the one with direct compositional stakes.

Quotas. A hard target for the disadvantaged share is the limiting case $G'' \to -\infty$: it predicts $\kappa = 0$ and no cross-reader dispersion in D_r. Both predictions fail: $\hat{\kappa} = -0.25$ is significantly negative, and reader-specific cohorts span a nontrivial range of disadvantaged shares (18 to 21 percent). University B's behavior is consistent with at most an elastic compositional preference, not a quota.

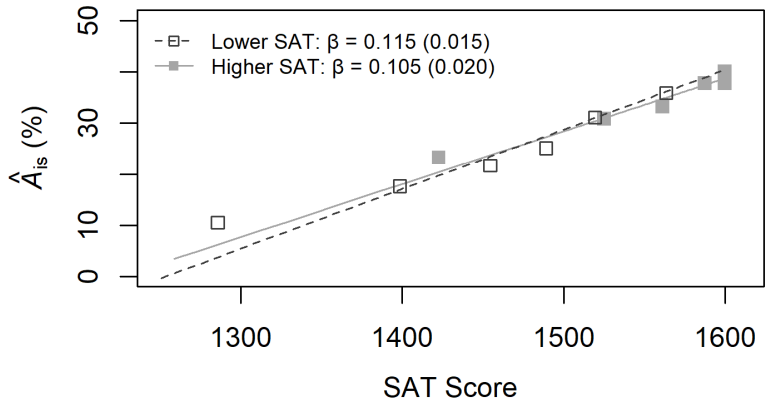
Sequential and yield-based mechanisms. Cohort-dependence could also arise from readers calibrating against recently-read applications or from admission decisions embedding enrollment-yield considerations. Our design largely forecloses both: packet composition is randomized (and cluster-by-day fixed effects absorb residual packet-level variation), and readers evaluate applications without yield information, so neither mechanism can generate the estimated gradient.

Summary

The applicant-level bonus λ and the cohort-composition gradient κ coincide only when the university's internal variation operates through the bonus itself. At University B it does not: the bonus is large but uniform across the realizations of the university's objective, while the variation that distinguishes those realizations — leniency and the relative weighting of academic against non-academic strength — is nearly orthogonal to disadvantage. The cohort-level price of disadvantaged enrollment is consequently set by the applicant pool's margin geometry ($b_m = -0.27$), roughly a third of the price a single-index extrapolation from the bonus would predict.

Other Appendix Figures and Tables

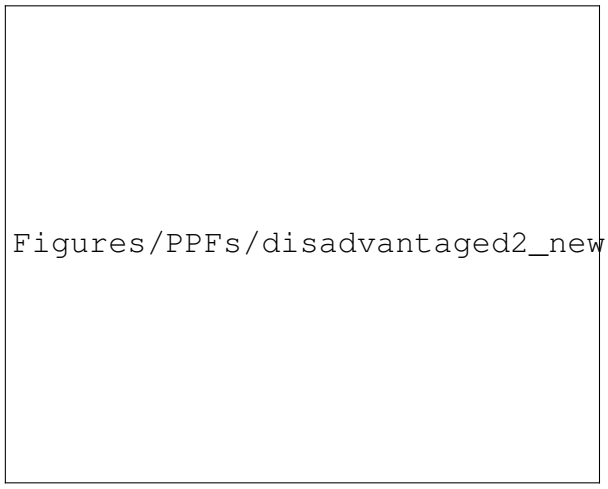
Figure A-1: The Causal Effect of Test Scores by the Direction of Experimental Manipulation



Note: The within-applicant relationship between SAT score and predicted admission likelihood ($\hat{A}_{isr}$) in the experimental sample split by whether the experimental test score is greater than or weakly less than the applicant's true test score, estimated with applicant fixed effects and scaled to the mean applicant. The legend includes coefficients (and standard errors in parentheses) of the slopes, which are not statistically distinguishable. Predicted admission ($\hat{A}_{isr}$) is the predicted likelihood of admission given the applicant's overall, academic, and non-academic ratings assuming they were evaluated by the average-leniency reader, as estimated using the second-round evaluations of second-round applicants excluded from the experiment. Standard errors are clustered by region-grouping (40 total clusters). Source: University B Administrative Data.

Figure A-2: Full PPFs

(a) Disadvantaged



(b) Gender



(c) Low SEI



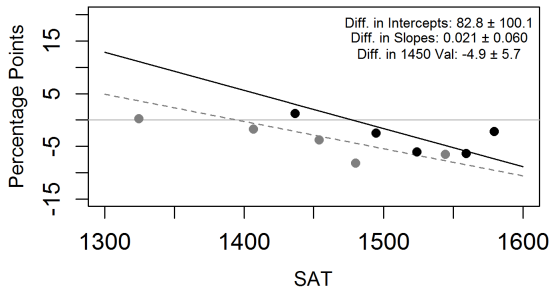
(d) URM



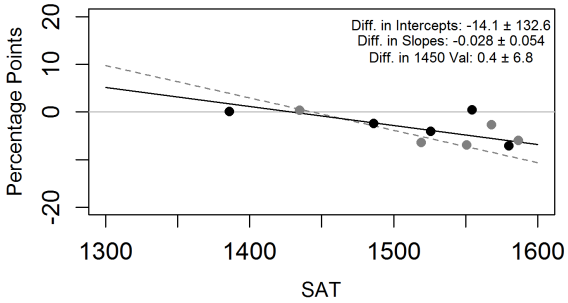
Note: XXX. Source: University B Administrative Data.

Figure A-3: Test Optional

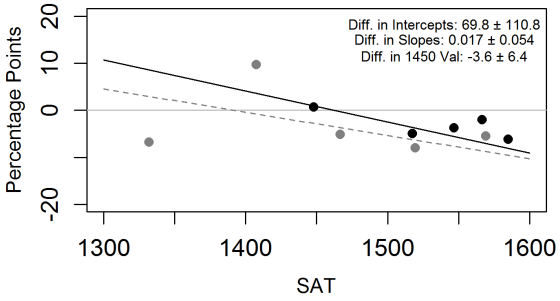
(a) Disadvantaged



(b) Gender

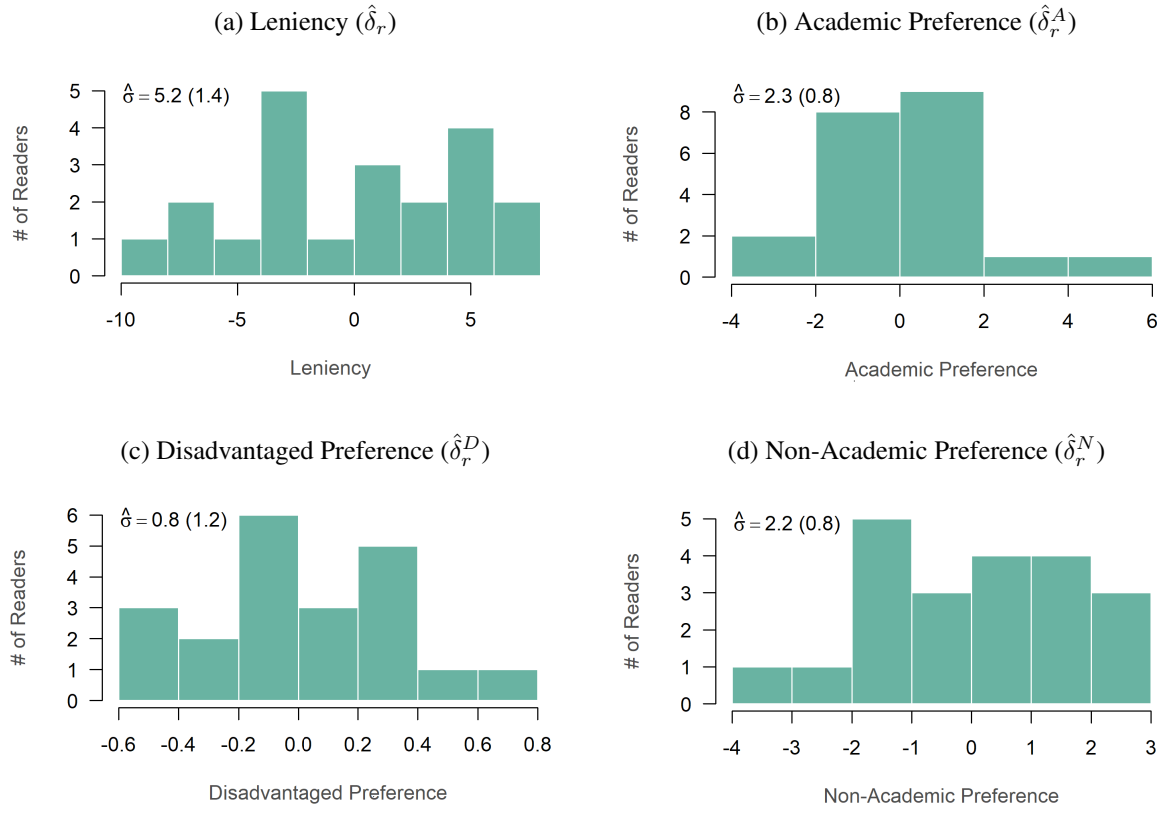


(c) URM



Note: XXX. Source: University B Administrative Data.

Figure A-4: Distribution of Reader Characteristics



Note: The shrunk distributions of reader leniency ($\hat{\delta}_r$) and readers' preferences for applicants with strong academics ($\hat{\delta}_r^A$), disadvantage ($\hat{\delta}_r^D$), or non-academics ($\hat{\delta}_r^N$) among the 21 admissions readers who participated in the audit study, and the estimated shrunken standard deviation of each. See Equation 4 for the definitions of each characteristic. Source: University B Administrative Data.

Table A-1: The Relationship between Admission and Applicant Characteristics at University B

	Admission to University B (%)							Disadvantaged	
Sample:	All Normal Applicants			Second-Round Applicants				Second-Round App.	
SAT (per 100)	2.5	3.1	3.6	9.4	13.5	11.4	10.0	-0.27	-0.27
	(0.1)	(0.1)	(0.2)	(1.3)	(1.4)	(2.1)	(2.2)	(3.0)	(2.9)
Disadvantaged		2.8	3.0		15.2	13.6	13.6		
		(0.3)	(0.3)		(2.4)	(2.4)	(2.5)		
Controls			X			X	X		
Weight							$\hat{A}(1-\hat{A})$		$\hat{A}(1-\hat{A})$
Adj. R^2	0.02	0.03	0.06	0.02	0.03	0.09	0.09	0.20	0.20
Observations	23,606	23,606	23,606	2,816	2,816	2,816	2,816	2,816	2,816
Implied $\hat{\lambda}$		91.9	85.6		112.6	119.5	136.7		
		(9.0)	(9.1)		(16.3)	(27.4)	(34.9)		

Note: OLS regression coefficients from models of admission to University B on SAT score (out of 1600), an indicator for the student's being disadvantaged, and control variables estimated over all normal or second-round 2024 applicants. Normal applicants exclude early-decision, foreign, athlete, and Questbridge applicants; applicants who applied to small specialized programs or who were flagged by university administration; and applicants missing key demographic information. 'Disadvantaged' is defined as being Black, Hispanic, or both first-generation and an Expected Family Contribution (EFC) below $15,000 (which approximately corresponds to a parental income of $110,000). Controls include indicators for legacy status, having taken an AP class, gender, having visited campus for a tour, intended major discipline; and third-order polynomials of high school GPA, high school rank, number of AP courses, the number of AP scores of 5, the number of AP scores of 4, the average AP score, and all available statistics characterizing the applicant's high school and neighborhood from College Board's Landscape tool. The implied $\hat{\lambda}$ is defined as 100 times the ratio of the disadvantaged coefficient to the SAT coefficient (with standard errors from the Delta method); see Section 3.3.

Source: University B Administrative Data.

Table A-2: Correlations between Reader Characteristics

		Reader Preference for:		
	Leniency ($\hat{\delta}_r$)	Acad. ($\hat{\delta}_r^A$)	Disad. ($\hat{\delta}_r^D$)	Non-Acad. ($\hat{\delta}_r^N$)
Leniency ($\hat{\delta}_r$)	1			
Academics ($\hat{\delta}_r^A$)	0.296	1		
Disadvantage ($\hat{\delta}_r^D$)	0.221	-0.004	1	
Non-Academics ($\hat{\delta}_r^N$)	0.086	0.521	-0.421	1

Note: The unadjusted pairwise correlations among the 21 admissions readers who participated in the audit study between reader leniency ($\hat{\delta}_r$) and readers' preferences for applicants with strong academics ($\hat{\delta}_r^A$), disadvantage ($\hat{\delta}_r^D$), or non-academics ($\hat{\delta}_r^N$). See Equation 4 for the definitions of each.

Source: University B Administrative Data.

Table A-3: Detailed Decomposition of Variance in Highly-Selective Holistic Admissions

Sample:	All Students					Disad.	Adv.
	$\hat{A}_{isr}$	$Rate_{is}$	$Acad_{is}$	$NonAcad_{is}$	$\hat{A}_{isr}$	$\hat{A}_{isr}$	$\hat{A}_{isr}$
Percent of Variance Explained:							
Applicant	31.0	31.2	21.3	40.3	30.7	30.3	31.7
Test Score	10.6	9.0	28.4	-0.1	6.6	15.5	8.4
Reader	4.5	6.4	1.7	9.8	4.4	5.7	5.3
Reader Prefs:							
Academic					0.7		
Disadvantage					0.7		
Non-Academic					0.7		
Covariances:							
Applicant-Score	-2.0	-2.8	12.4	-0.3	0.6	-5.7	-0.6
Applicant-Reader	1.1	0.9	0.0	-0.5	1.0	-1.4	1.6
Score-Reader	0.3	0.6	0.0	0.0	0.3	1.0	0.6
Other[1]	0.5	0.6	0.0	0.2	2.5	0.6	0.5
Residual	53.3	53.8	36.3	50.5	51.3	52.7	52.2
# of Students			516			130	386
Observations			2,713			711	1999

Note: This table extends Table 2 by presenting decompositions of the variance of predicted admission ($\hat{A}_{isr}$) and the overall, academic, and non-academic rating of experimental 2024 University B applicant evaluations into shares explained by the full application portfolio (excluding the SAT score), the experimentally-manipulated SAT score, the assigned application reader, covariances between these terms (which includes the main effect and covariances with whether the evaluation occurred on the third test-mandatory day of the experiment), and the residual, including the broken-out reader preference and covariance components. Variance components are estimated using the leave-out estimator of Kline et al. (2020), which replaces each observation's squared-error contribution with an unbiased leave-one-out estimate; four observations with leverage 1 are excluded. The fifth column further includes terms capturing the assigned reader's relative preferences for disadvantaged students, strong academic students and applicants with strong non-academic credentials; the sixth and seventh columns restrict the decomposition to disadvantaged and non-disadvantaged applicants, respectively. 'Disadvantaged' is defined as being Black, Hispanic, or both first-generation and an Expected Family Contribution (EFC) below $15,000 (which approximately corresponds to a parental income of $110,000); predicted admission ($\hat{A}_{isr}$) is the predicted likelihood of admission given the applicant's overall, academic, and non-academic ratings assuming they were evaluated by the average-leniency reader, as estimated using the second-round evaluations of second-round applicants excluded from the experiment. [1]**Other covariances** include the variance attributable to an indicator for the evaluation's being conducted on the third day (with a test-mandatory regime), covariances between that indicator and the other covariates, and (in column 5) covariances between the three reader preference characteristics and all other covariates.

Source: University B Administrative Data.

Table A-4: Cohort-Level SAT–Disadvantage Gradients: Benchmark and Estimates

	Single-index benchmark	Cohort-composition gradient	Margin pool gradient
	$-100/\hat{\lambda}_m$	$\hat{\kappa}$	$\hat{b}_m$
Slope	-0.73	-0.25	-0.27
	(0.19)	(0.04)	(0.03)
Test-mandatory regime		-0.28	-0.27
		(0.05)	(0.03)
Bootstrap test: $\hat{\kappa}$ equals column slope (one-sided p)			
	0.01	—	0.25

Note: All statistics are slopes measuring the change in the disadvantaged share of an admitted cohort (in percentage points) per one-point increase in the cohort's mean SAT score. The single-index benchmark is the gradient implied by the margin-weighted applicant-level bonus $\hat{\lambda}_m = 137$ (Table A-1): under admissions that value disadvantage at a fixed SAT-point bonus, cohorts trade disadvantaged share against mean SAT at rate $-100/\lambda$ (Appendix C, Proposition 1); standard error by the Delta method. The cohort-composition gradient $\hat{\kappa}$ is the cross-reader slope from Figure 2; the margin pool gradient $\hat{b}_m$ is the $\hat{A}(1-\hat{A})$-weighted regression of disadvantage on SAT among second-round applicants (Table A-1). The test-mandatory row re-estimates each under the test-mandatory regime (Figure **??**b). The bottom row reports one-sided p-values from XXX cluster-bootstrap replications of the full estimation pipeline (resampling 40 region-groupings, re-estimating reader characteristics, shrinkage, and all slopes jointly), testing $\hat{\kappa}$ against each reference slope. Standard errors clustered by region-grouping. Source: University B Administrative Data.