Reassessing the Role of Standardized Tests in University Admissions*

Johann D. Gaebler[1], Calvin Isley[2], Christopher Avery[2], and Sharad Goel[2]

[1]Department of Statistics, Harvard University
[2]John F. Kennedy School of Government, Harvard University

April 7, 2026

Abstract

There is a long-running debate over using standardized test scores to inform college and graduate admissions decisions, with some arguing that test scores are an important signal of academic qualification and others arguing that they increase inequality. Here we revisit this issue by analyzing a novel dataset of more than 13,000 applications over roughly a decade to a large public policy master's program in the United States. Consistent with past work, we find that GRE scores substantially improve predictions of first-year grades, relative to predictions based on GPA alone. However, when these predictions are used to inform admissions decisions, we find that test scores only modestly improve the expected academic quality of admitted students. The gap shrinks further when we augment the test-aware and test-blind predictive models with more fine-grained information available in student transcripts and other application materials. This pattern stems from a subtle distinction between predictions and decisions. Even with improved predictions, the downstream admissions decisions are often the same; and where there are differences, they often involve selecting between similarly qualified applicants. Our results indicate that standardized test scores may be less important for university admissions than previously suggested.

*The authors thank Deirdre Bloome, Jennifer Hochschild, Daniel Schneider, Maya Sen, David Weil, and the participants in the Stone Inequality Proseminar for helpful discussions. The authors also thank the admissions team at the participating institution for guidance and data access.

1 Introduction

Standardized testing has been central to American higher education for the past half century. By the mid-1970s, millions of prospective college applicants took the SAT and ACT each year (College Entrance Examination Board, 1975). Despite the ubiquity of standardized testing, there is a longstanding debate on the appropriateness of testing requirements. Proponents often argue that standardized assessments provide crucial information about an applicant's academic preparation not otherwise available to admissions officers (Friedman et al., 2025). Critics, however, counter that these assessments reinforce obstacles facing groups underrepresented in higher education (Rosinger et al., 2021). They object to potential biases in the tests themselves (Goodman et al., 2020), differential access to test preparation services (Buchmann et al., 2010), and weak predictive validity for key educational outcomes (Feldon et al., 2024). Testing proponents have responded that eliminating testing requirements has largely harmed rather than helped disadvantaged student groups (Sacerdote et al., 2025).

Opposition to standardized testing peaked in recent years, coinciding with the onset of the global COVID pandemic in 2019. In response, many colleges and universities initially relaxed or eliminated testing requirements to ease logistical burdens for students. When the pandemic subsided, many institutions chose not to bring back the requirements. In particular, during the 2023–2024 admissions cycle, less than half of Common App applicants reported a standardized test score, compared to roughly 75% in 2019; and the proportion of Common App member schools requiring test scores fell from 55% to 5% over the same period (Kim et al., 2024). However, there is now renewed support for standardized testing, spurred by new empirical evidence highlighting its benefits (e.g., Cascio et al., 2024; Harvard University, 2024). Seven of the eight Ivy League universities have now either reinstated testing requirements or plan to do so in the upcoming admissions cycle.

The seesawing policies and competing claims over testing reflect a fundamentally unresolved question: How much value do standardized test scores add to admissions? Here we provide a new theoretical frame for answering this question. We ground our analysis with a novel dataset consisting of more than 13,000 applications to a large public policy master's program at a competitive U.S. professional school. In predicting future GPA, we find that statistical models leveraging standardized test scores substantially outperform those that only consider past GPA—consistent with the existing literature (Cascio et al., 2024; Chetty et al., 2023; Friedman et al., 2025; Kuncel and Hezlett, 2007; Kuncel et al., 2001; Rothstein, 2004). However, the problem facing admissions officers is not to best *predict* future academic performance, but rather to optimally *select* high-achieving students. While related, the prediction problem commonly considered in past empirical analyses of standardized testing is distinct from the optimization problem that academic institutions must in reality solve.

Reframed in this way, we find that standardized test scores yield relatively small improvements to the quality of admissions decisions. When decisions are based on past GPA and other common baseline covariates, we estimate that the admitted class achieves an average GPA of 3.58 in the required first-semester economics and statistics courses; if GRE scores are additionally considered, the average GPA of admitted students increases modestly to 3.63. Even though predictions shift, the decisions themselves are relatively stable. In other words, it is easier to identify top performers than to predict exactly how well they ultimately

fare. Further, to the extent that decisions do change, they typically involve rejecting or accepting similarly qualified applicants who are close to the decision boundary. As a result, while test scores substantially improve predictive quality, their impact on decision quality is more limited.

The marginal value of test scores further attenuates when we supplement our baseline covariates with detailed information drawn from applicants' undergraduate and previous graduate transcripts, letters of recommendation, resumes, and application essays. To incorporate this additional data, we leverage modern AI-enabled document parsing pipelines, which can reliably extract structured information from these materials, ranging from course information in transcripts to sector-specific work experience in resumes to measures of endorsement strength in recommendation letters. After doing so, the added decision-relevant value of standardized test scores shrinks by about half, yielding an admitted class that, on average, performs better by only 0.03 grade points.

Our findings highlight the importance of distinguishing between predictions and decisions, as well as incorporating the full range of information applicants provide, when evaluating the marginal value of standardized test scores for admissions. Our empirical analysis focuses on a single graduate policy program at a selective institution. Though our general statistical observations apply more broadly, the precise magnitude of the effects we estimate likely vary—perhaps substantially—across settings.

2 Data and Methods

We base our analysis on the complete set of applications to a two-year public policy master's program at a competitive American professional school between the 2013–2014 and 2023–2024 admissions cycles. The program implemented a test-optional admissions policy during the 2020–2021 and 2021–2022 admissions cycles, which we drop from our analysis, leaving nine complete rounds and 13,092 distinct applications that we analyze. We link these applications to the academic performance of the 2,159 students who ultimately attended the program.

These applicants comprise a broad range of professional and demographic backgrounds, including international applicants who make up about 40% of the pool. Admission to the professional school is based on a holistic review of a variety of factors, including test scores, academic performance in undergraduate and previous graduate study, work experience, letters of recommendation, and personal essays. Enrolled students are required to take a battery of standard courses in their first year, followed by elective courses in their second year.

Covariates

We observe prospective students' complete applications to the public policy program. In line with existing literature (Friedman et al., 2025), we take students' college GPAs and their GRE scores as our focal covariates. We supplement these two measures with a collection of "baseline covariates" capturing information about the strength of applicants' undergraduate institutions. To do so, we draw from two sources: (1) the 2025 QS World University Rankings (QS Quacquarelli Symonds, 2024), which rank approximately 1,500 universities

worldwide on a variety of dimensions; and (2) IPEDS survey data (National Center for Education Statistics, 2024), which gather comprehensive institutional characteristics for virtually all U.S. postsecondary institutions. Following the 2023 Supreme Court decision in *SFFA v. Harvard*, legally protected characteristics such as race generally cannot be lawfully considered in admissions in the U.S. (SFFA v. Harvard, 2023). We therefore exclude from consideration race, gender, age, nationality, and other demographics applicants report, as admissions policies that explicitly consider these factors likely could not be put into practice.

Admissions officers typically consider standardized test scores and GPAs against the backdrop of a wide variety of other potentially relevant information in applicants' submitted materials. We correspondingly extract a detailed collection of "comprehensive covariates" from academic transcripts, letters of recommendation, resumes, and application essays. We use Microsoft Document Intelligence (Microsoft, 2024) to extract the text contents of the documents, in some cases after more specialized preprocessing. We then use various OpenAI language models (OpenAI, 2025) to structure the raw text. In particular, we use the language models both to extract explicitly provided information (e.g., years of work experience and course grades and titles), and also to evaluate applicants along several dimensions based on structured rubrics (e.g., adherence of essays to the provided prompt and recommenders' level of endorsement of applicants' quantitative analysis skills). In total, we extract 396 detailed measures from these materials (e.g., whether the applicant took an undergraduate microeconomics class and, if so, the grade they received). In the Appendix, we additionally consider models incorporating only covariates from transcripts and models which do not incorporate any covariates beyond GRE scores and college GPAs. See Appendix A.1 for full details on the data cleaning and featurization process.

Outcomes

At the public policy program we consider, an important challenge in the admissions process—and a primary motivation for requiring standardized tests—is ensuring that admitted students are adequately prepared for the required first-semester statistics and economics courses. These courses have relatively stable curricula and grade distributions (Fig. 1), making them an informative measure of academic strength. We take a student's average performance in these two courses, measured on the 4.0 scale, as our primary outcome of interest. In the Appendix, we also consider a variety of other outcomes, including students' performance across the entire set of quantitative first-year courses; performance across all courses taken in their first year; course passage; and a holistic composite rating, grounded in actual admissions criteria and decisions, incorporating measures of leadership, interest in public policy, and other factors beyond academic preparation. We find qualitatively similar results across these different outcome measures. See Appendix A.2 for full details.

A key statistical challenge in studying admissions policies is that grades and other outcomes are generally only available for applicants who are admitted and attend. To account for this, we impute outcomes for non-matriculating applicants using the broadest available range of covariates (Rubin, 1987), including demographics, basing our subsequent analyses on the imputed results. Using these covariates, we can predict admissions decisions very accurately, achieving an AUC of 89% [CI: 84%–93%], suggesting limited scope for admitted and rejected applicants to differ in unobserved ways. However, if applicants who do

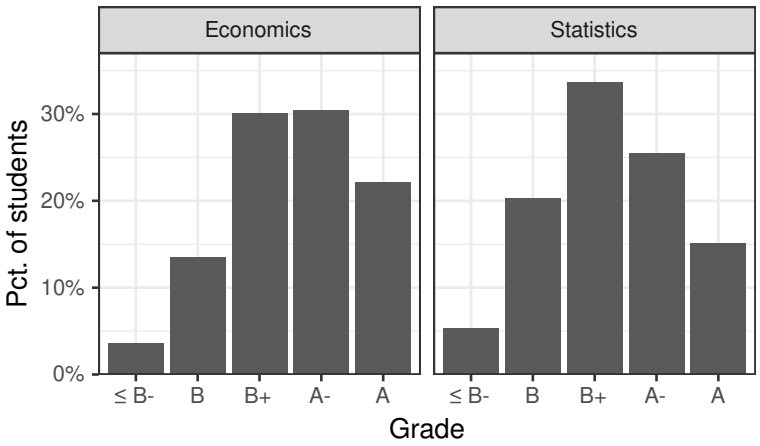

Figure 1: *The distribution of grades in the required first-semester economics and statistics courses. Very few students receive grades lower than a B-, and most students in both courses receive a B+ or A-, while only 15% receive an A in statistics and 22% in economics.*

not enroll—and hence whose grades we do not observe—differ from the enrolled students in unobservable ways, our estimates may suffer from omitted variable bias. In the Appendix, following existing literature in this area (Friedman et al., 2025), we repeat our analyses using only enrolled students, for whom outcomes are observed directly and do not need to be imputed. We additionally consider the subset of applicants with at least a 10% *ex ante* chance of being admitted and enrolling in the masters program, for whom the imputation models are likely most accurate. In all cases, we find qualitatively similar results. See Appendix A.3 for full details on these different approaches.

Modeling

Our primary analyses use gradient-boosted decision trees—a popular non-linear regression method in the machine learning community—fit using the `xgboost R` package (Chen and Guestrin, 2016). We generate out-of-sample predictions for each of the nine rounds of admissions we consider by training a model to predict academic performance on the remaining eight rounds, tuning model hyperparameters using nested cross-validation. We repeat all analyses using both penalized and unpenalized linear models and find qualitatively similar results; see Appendix A.3. We report cross-validated estimates of model performance, admitted class academic and demographic outcomes, and other quantities. Our confidence intervals account for multiple sources of uncertainty: year-to-year variation in the applicant pool, randomness in the model tuning and fitting process, and outcome missingness through multiple imputation (Rubin, 1987).

3 Results

We start by assessing the accuracy of statistical models that predict academic performance both with and without test scores. We then investigate how the quality of a model's predictions connects to the quality of the downstream decisions it facilitates—first empirically

via policy simulations (Grossman et al., 2024) and then theoretically. Finally, we strengthen the overall quality of our models by incorporating detailed information from transcripts and other application materials. We primarily focus on the overall academic strength of the class of students that can be selected with and without using test scores. However, we also study how incorporating test scores affects the full grade distribution and demographic composition of selected students. In particular, we consider whether test scores help one identify academically talented students from less selective colleges who might otherwise be passed over, as some have argued.

Predictive performance

Mirroring past analyses (Sacerdote et al., 2025), Figure 2 shows that (quantitative) GRE scores are a strong predictor of grades in first-semester statistics and economics courses, even after accounting for a student's undergraduate GPA. For example, after adjusting for past GPA, every 4-point increase in quantitative GRE score corresponds to approximately a 0.1 point increase in expected GPA in first-semester quantitative classes. A model that includes only past GPA achieves an R^2 of 8% [CI: 6%–9%], whereas adding in GRE scores boosts it to 31% [CI: 27%–34%], a four-fold improvement.[1] This dramatic increase in predictive quality is in line with past results, and is the primary empirical fact driving claims that standardized tests are critically important for informing admissions decisions (Friedman et al., 2025).

One reason that past GPA is not more predictive is that grades often mean different things at different institutions. To account for this fact, we augment our two models above—those with and without GRE scores—with information on the strength of applicants' undergraduate institutions, as described in Section 2. The resulting "baseline" model without GRE scores achieves an R^2 of 18% [CI: 16%–21%], whereas the corresponding model with GRE scores achieves an R^2 of 35% [CI: 31%–38%], still approximately doubling the GRE-blind model's predictive accuracy. In sum, whether or not we include additional baseline covariates, we find that standardized tests substantially improve predictions of academic performance over GPA alone, replicating past results.

Predictions vs. decisions

The predictive quality of statistical models is only indirectly connected to the quality of the downstream decisions they facilitate (Coots et al., 2025). To study this latter decision problem more directly, we next investigate the academic quality of the top students selected via statistical models that include or omit GRE scores. Specifically, we begin by ranking applicants according to their predicted first-semester quantitative GPA under two models:

[1] In OLS regression, R^2 is equivalently defined as: (1) the squared correlation between predictions and the dependent variable, or (2) one minus the fraction of variance unexplained. In general, however, these two measures need not be equal. Because of its connection to Proposition 1, we adopt the former definition throughout. To correct for range restriction, we adjust our R^2 values to reflect predictive performance over the full set of applicants, using imputed grades for those applicants who did not ultimately attend the institution; cf. Rothstein Rothstein (2004). The unadjusted R^2 values—computed only over matriculants—show the same general pattern, with an R^2 of 5% for the model without GRE scores and 25% for the model with GRE scores. All confidence intervals shown are 95% confidence intervals; see Appendix A.3 for full details.

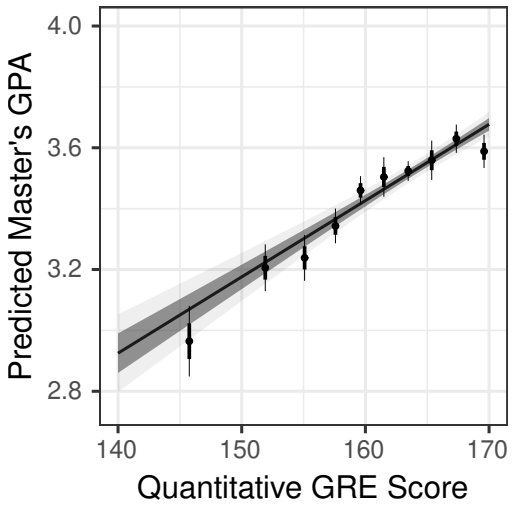

Figure 2: *The expected first-semester quantitative GPA as a function of quantitative GRE score, after adjusting for past undergraduate and graduate GPA. The line shows estimated first-semester quantitative GPA as a linear function, with points indicating fixed effects estimated when GRE scores are instead binned into 10 equally sized bins. Dark shaded ribbons and thick lines indicate 68% pointwise confidence intervals; light shaded ribbons and thin lines indicate 95% pointwise confidence intervals.*

(1) a baseline model that includes past GPA and measures of school strength, and (2) a test-aware model that additionally includes GRE scores. To mimic the admissions constraints of a selective institution, we then select the top 20% of applicants under each model, simulating "GRE-blind" and "GRE-aware" admissions policies. (We also consider more and less selective admissions policies and find qualitatively similar results; see the Appendix.)

Under the GRE-aware admissions policy, the admitted class achieves an average GPA of 3.63 [CI: 3.60–3.66], 0.051 [CI: 0.031–0.070] grade points higher than under the GRE-blind policy. While non-negligible, the improvement is perhaps smaller than expected given that the R^2 of the GRE-aware model is double that of the GRE-blind model. Going beyond average GPA, Figure 23 in the Appendix plots the full estimated grade distribution under the two policies, and shows that the outcomes are again surprisingly similar across the distribution.

A decision-theoretic framing helps explain this result. Admissions officers face a binary choice—whether to admit or reject each applicant. Predictions of future academic performance inform that choice, but the decision problem differs fundamentally from the task of prediction itself. This distinction is illustrated in Figure 3. The horizontal axis shows predicted first-semester grades under the GRE-aware model, with the dashed vertical line indicating the admissions threshold under that model. That is, the GRE-aware policy admits the 20% of applicants to the right of the dashed line. In contrast, the GRE-blind model admits a different set of students, indicated in blue, some of whom are below the dashed line. In the absence of test scores, the GRE-blind model ranks those students in its own top 20%.

The figure illustrates two key points. First, while the GRE-aware and GRE-blind models

Table 1: Comparison of GRE-Aware and GRE-Blind Admissions Policies

	Without GRE	With GRE	Diff.
Baseline			
Range Adj. R^2	0.183^{***} (0.160, 0.206)	0.349^{***} (0.314, 0.384)	0.166^{***} (0.139, 0.193)
Range Unadj. R^2	0.160^{***} (0.118, 0.202)	0.280^{***} (0.226, 0.334)	0.120^{***} (0.075, 0.165)
Class GPA	3.58^{***} (3.54, 3.62)	3.63^{***} (3.60, 3.66)	0.051^{***} (0.031, 0.070)
Pct. Decisions Differing			$13.4\%^{***}$ (12.4%, 14.3%)
Comprehensive			
Range Adj. R^2	0.239^{***} (0.201, 0.277)	0.350^{***} (0.308, 0.392)	0.111^{***} (0.088, 0.135)
Range Unadj. R^2	0.248^{***} (0.189, 0.308)	0.324^{***} (0.262, 0.385)	0.075^{***} (0.045, 0.106)
Class GPA	3.62^{***} (3.58, 3.65)	3.64^{***} (3.61, 3.67)	0.028^{***} (0.017, 0.040)
Pct. Decisions Differing			$9.5\%^{***}$ (8.7%, 10.4%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted R^2 values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted R^2 values are computed over matriculants without imputation. We report out-of-sample values for both measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

produce different predictions for each applicant, the discrepancy between the two predictions is typically too small to change the admissions decision. Nearly 90% of applicants are either accepted or rejected regardless of which model is used to inform that decision. Second, when decisions do differ between the two models, they tend to involve applicants near the decision threshold, who are academically similar to one another. Consequently, the net effect of these students' entering and leaving the cohort of admitted students—that is, of exchanging the blue portion of the distribution to the left of the dashed line with the gray portion to the right of the dashed line—on the overall distribution of grades is necessarily small.

A simple statistical heuristic sheds light on this empirical pattern. Let $\varphi(z)$ and $\Phi^{-1}(p)$ denote the standard normal PDF and inverse CDF, respectively. Then, the following proposition is a straightforward consequence of well-known probabilistic facts.

Proposition 1. *Let Y, $\hat{Y}_0$, and $\hat{Y}_1$ be jointly normal random variables. Then*

$$\mathbb{E}[Y \mid \hat{Y}_1 > y_1] - \mathbb{E}[Y \mid \hat{Y}_0 > y_0]$$

equals

$$\sigma \cdot (\rho_1 - \rho_0) \cdot \frac{\varphi(\Phi^{-1}(p))}{p}, \tag{1}$$

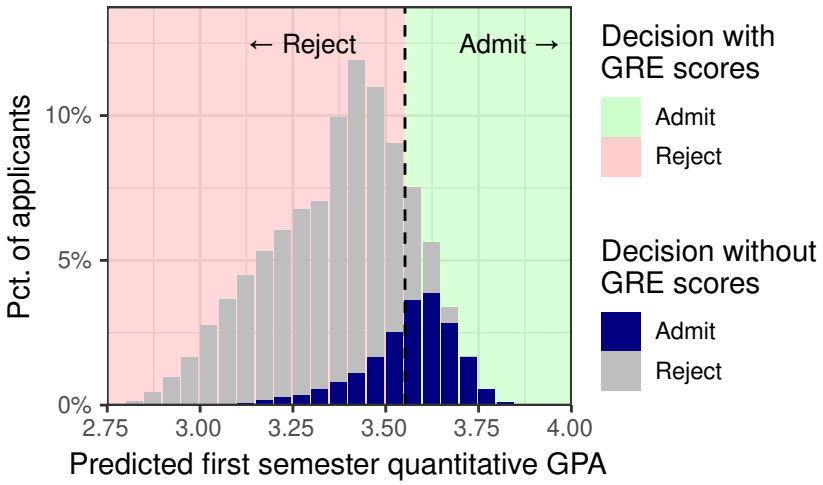

Figure 3: *The distribution of applicants' predicted first-semester quantitative GPA under the GRE-aware model, where color indicates whether the applicant is rejected (gray) or accepted (blue) under the simulated GRE-blind admissions policy and the dashed vertical line indicates the decision threshold for the GRE-aware admissions policy.*

with $\rho_i \stackrel{\text{def}}{=} \textsc{Cor}(Y, \hat{Y}_i)$, $\sigma^2 \stackrel{\text{def}}{=} \textsc{Var}(Y)$, *and* y_i *the* $(1-p)$*-quantile of* $\hat{Y}_i$, *i.e.,* $\Pr(\hat{Y}_i > y_i) = p$.

In our setting, Y denotes a student's realized first-semester GPA, and $\hat{Y}_0$ and $\hat{Y}_1$ their predicted GPA under the GRE-blind and GRE-aware models, respectively. Thus, $E[Y \mid \hat{Y}_1 > y_1] - \mathbb{E}[Y \mid \hat{Y}_0 > y_0]$ is the expected difference between the average GPA of the top p-percent of students admitted under the two models. The proposition shows that this difference is a product of three terms. The first term, σ, is a scaling factor that captures the underlying variance of the outcome. The second term, $\rho_1 - \rho_0$, encodes the difference in the models' predictive performance, since by definition ρ_i^2 is the R^2 of the respective models. Finally, the third term is a function of the selection probability p, which, for example, equals 1.4 for $p = 0.2$ and 2.1 for $p = 0.05$. The key substantive assumption is that the predicted and actual values are jointly normal, which holds approximately, though imperfectly, in our data; see Figure 7 in the Appendix.

Figure 4 compares the theoretical predictions of Eq. (1) to the actual differences in GPA we observe in our policy simulations across the full range of admission rates. We find that the theoretical and empirical estimates match well over the full spectrum. In particular, the theoretical estimates correctly suggest that even large differences in predictive quality translate to modest differences in the quality of the admitted cohorts.

Incorporating comprehensive application information

Our analysis thus far has focused on relatively parsimonious predictive models, which include past GPA, measures of undergraduate school strength, and GRE scores. That approach is standard in the literature, but fails to account for the wide variety of information applicants submit and admissions officers consider, from detailed college transcripts, to resumes, to letters of recommendation and personal essays. Modern AI-enabled data processing pipelines

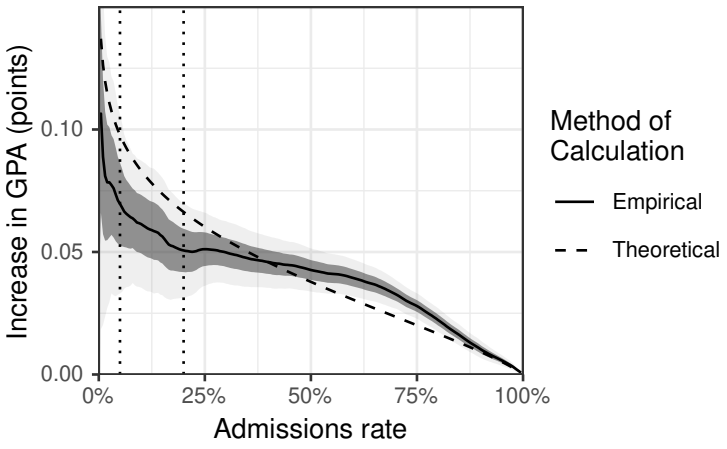

Figure 4: *A comparison of the increase in GPA in the admitted class resulting from including standardized test scores in the prediction model across admission rates. The solid and dashed lines represent whether the difference is based on our policy simulation or* (1)*, respectively. The dotted lines indicate a selective (20%) admissions policy as well as a highly selective (5%) admissions policy, matching the admissions rates at the most competitive U.S. undergraduate institutions (National Center for Education Statistics, 2024). Dark and light shaded ribbons indicate 68% and 95% pointwise confidence intervals, respectively.*

can now reliably extract this information from diverse and inconsistently formatted application materials—including, for example, course titles and grades appearing on transcripts. One would expect this additional data to at least partially offset the informational value of standardized test scores.

To test this hypothesis, we reproduce the foregoing analyses, adjusting instead for the comprehensive covariates described in Section 2 above. With the comprehensive covariates included, we find that the R^2 of the GRE-blind model increases to 24% [CI: 20%–27%]. The gain is negligible for the GRE-aware model, shrinking the gap in predictive performance to 11 p.p. [CI: 9 p.p.–13 p.p.]. Incorporating comprehensive covariates likewise raises the overall academic quality of the admitted cohorts and narrows the gap between GRE-aware and GRE-blind policies. Specifically, we find a difference of 0.028 [CI: 0.017–0.040] grade points between GRE-aware and GRE-blind admissions, just over half as large as before. In absolute terms, the average quality of admitted students under the two policies is quite similar.

We have so far considered the average GPA of selected students as a proxy for the overall quality of the admitted class. However, beyond this average, institutions may care about the distribution of outcomes across admitted students—particularly the proportion of low-performing students, who may require additional resources to succeed. Figure 5 plots the full distribution of estimated grades in the first-semester economics and statistics courses under the GRE-blind and GRE-aware models (with comprehensive covariates). We find that the differences between the two admissions policies are quite modest across the grade distribution.

Relatedly, considering or ignoring test scores may change the demographic composition of admitted students. GRE-blind or, conversely, GRE-aware policies may inadvertently disadvantage certain groups. For instance, test scores may provide an avenue for highly

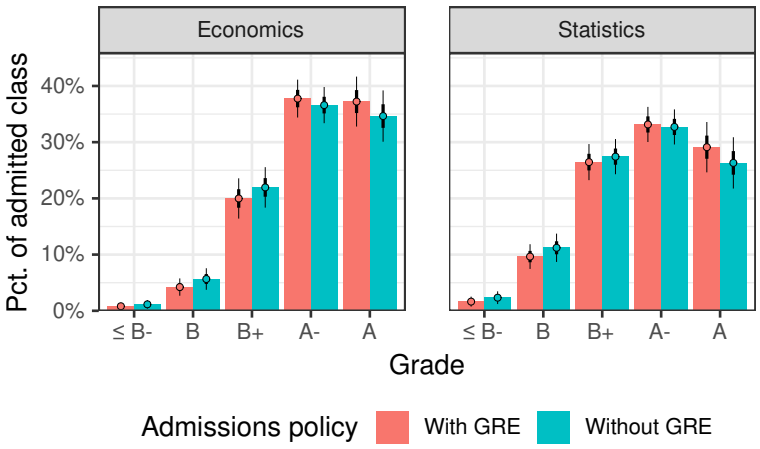

Figure 5: *The distribution of grades in (simulated) cohorts of students admitted according to their predicted academic performance both with (red) and without (blue) standardized test scores. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

qualified students from low-ranking academic institutions to better signal their preparation for selective educational programs. As a result, fewer students from low-ranking institutions may be admitted under GRE-blind policies. Figure 6 plots the demographics composition of students selected by GRE-aware and GRE-blind models. Along the four dimensions we consider—rank of undergraduate institution, race, gender, and international status—we find that the two policies yield nearly identical results. The largest gap is for gender, where GRE-aware policies result in admitting slightly more men. When decisions are nearly identical between GRE-aware and GRE-blind policies, there is little room for the cohorts of admitted students to differ much along any dimension, academic or otherwise. We formalize this intuition in Appendix B.2.

Robustness checks

We reproduced our main analysis across the full range of five outcomes (first-semester quantitative grades, first-year quantitative grades, first-year overall grades, course passage, and holistic admissions criteria) along with two additional variations of quantitative grades described in Appendix A.1, four covariate sets (minimal, baseline, transcript-only, comprehensive), three subpopulations (all applicants, enrolled applicants, "overlapping" applicants), three levels of selectivity (10% admissions rate, 20% admissions rate, 30% admissions rate), and model types (unpenalized linear regression, penalized linear regression, ensembled GBMs). Across specifications, we see results broadly consistent with the following findings: (1) models incorporating GRE scores yield better predictions than those without them, typically by a large margin; (2) academic and demographic outcomes among applicants selected by these predictive models exhibit more modest gaps than performance differences would suggest; and (3) adjusting for more complete information about applicants improves performance and reduces or eliminates outcome gaps.

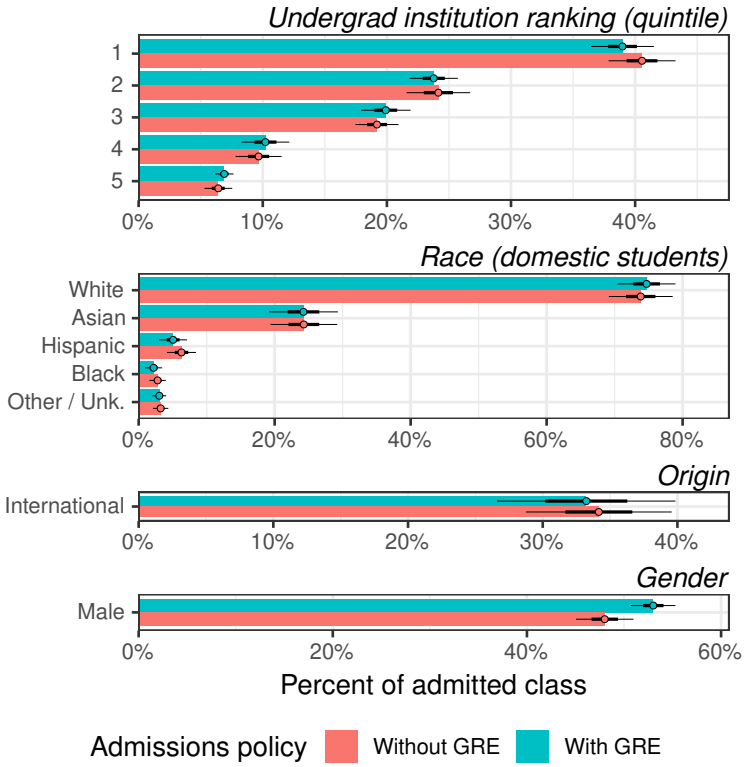

Figure 6: *A comparison of the demographics of students admitted under the GRE-aware and GRE-blind admissions policies using comprehensive covariates. The facets show the distribution of admitted students' undergraduate institutional rank, self-reported race (domestic students only), origin, and gender. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

4 Discussion

Recent empirical analyses have concluded that standardized tests are a critical factor for university admissions, prompting many institutions to reinstate standardized testing requirements. Our work, however, raises two statistical concerns with that conclusion. First, while past work has focused on the *predictive* value of test scores, more relevant is the extent to which test scores improve admissions *decisions*—a subtle but important distinction. Second, it is common to assess the marginal value of test scores against a relatively weak baseline comprised only of past GPA and basic demographics. In reality, admissions officers consider a much richer set of information that can now also be incorporated into statistical models. After addressing these two concerns, we find that test scores yield only slight improvement to the academic quality of selected students, at least in our setting. Moreover, we find that the demographics of students selected under test-blind and test-aware predictive models are nearly identical, allaying concerns that considering or ignoring test scores may harm certain groups.

When interpreting our results, several important limitations bear emphasis. First, and most critically, our findings may only partially generalize beyond the specific context we study. In particular, it is possible that standardized test scores are relatively more informa-

tive for college applicants, as there are many high-achieving high school students who earn near-perfect grades, making it harder to distinguish between them without test scores. On the other hand, high school transcripts contain a plethora of information beyond average GPA—including which specific math, science, and AP courses a student has taken—which likely offsets at least some of the predictive power of standardized test scores. Predictions may also improve by accounting for the myriad activities that high school students often engage in, from sports to school clubs to volunteering and part-time jobs. Relatedly, Proposition 1 suggests that gaps in grade outcomes increase as admissions rates drop. Consequently, test scores may be more important at the highly selective colleges that are often the focus of testing debates. Ultimately, more empirical evidence is necessary to determine in which settings, and exactly how much, test scores matter for admissions decisions.

Second, in our primary analysis, we impute outcomes for the large majority of applicants who never attended the public policy program we consider. To address this concern, we perform the imputation using rich information about applicants, incorporating demographic covariates along with covariates summarizing application materials, to reduce the severity of potential selection bias. To change our findings, unobserved confounders not only have to affect the performance of our imputation models, but to do so in a way that differs substantially when standardized test scores are or are not included as predictors. We additionally replicate our analyses in two ways: (1) on enrolled students, for whom imputation is not necessary; and (2) on the subset of applicants observably similar to enrolled students, for whom imputation is likely most reliable. Both robustness checks produce results broadly consistent with our main analysis.

Third, we built a complex data processing pipeline to extract the rich information in transcripts, letters of recommendation, resumes, and application essays; that approach may be prohibitively difficult for some institutions to implement. By comparison, standardized test scores are a logistically simple measure to incorporate into predictive models. Relatedly, our approach is predicated on using statistical models to inform decisions, which can efficiently extract signals from complex data. Human decision makers, unaided by statistical tools, may struggle to reliably assess applicants without standardized test scores. Without appropriately training admissions officers, removing testing requirements could thus cause them to admit less academically prepared cohorts (Senate-Administration Working Group on Admissions, 2025).

Fourth, we have focused on assembling the most academically prepared class possible. However, most institutions—including the program we study—pursue other academic and non-academic objectives. For instance, beyond academic preparation, policy programs often favor applicants likely to pursue government or public-interest work. The purely academic scope of most standardized testing suggests that introducing competing objectives would further shrink the discrepancies between test-blind and test-aware admissions policies. Indeed, when we consider a holistic admissions objective that accounts for public service, leadership, and other qualities the program we study seeks in its students, we observe exactly this pattern; see Appendix C.[2]

[2]We first assign each matriculated student a score based on a variety of academic and non-academic factors according to rubrics provided to us by the admissions office. These scores include, for instance, students' first-semester quantitative grades and public service experience. Then, as before, we predict these scores based on statistical models that include or omit test scores. Finally, we simulate test-aware and test-

Finally, our results reflect the current environment in which standardized testing remains a significant admissions factor. Eliminating testing requirements may change the pool of applicants in important ways (Avery et al., 2025; Garg et al., 2026). Further, if transcripts became more heavily weighted in admissions, for instance, students might respond by enrolling in institutions with more lenient standards, thereby reducing the predictive power of transcripts and other detailed application information. Similar unanticipated consequences could arise if students change their study habits in response to more lenient testing requirements. These and other equilibrium effects are difficult to estimate with our data and methods.

Returning to our motivating question, should colleges and universities ultimately eliminate or maintain standardized testing requirements for admissions? Our findings suggest that eliminating testing requirements would have limited effects on the academic quality of admitted students—at least when decisions are based on statistical models that consider the full set of materials that applicants submit. Standardized testing may, however, have additional benefits, like offering a simple, transparent measure for ranking students. Further, we have not assessed the costs that these tests impose on institutions and students. Even if the benefits of standardized testing are small, the costs may be equally small or smaller, leaving open the possibility that testing requirements are justified. Rather than providing a definitive answer to the testing question, our hope is to reframe the terms of the debate. By shifting the focus from predictions to decisions, we aim to equip researchers and university administrators with a more rigorous, statistically grounded foundation for assessing admissions practices.

blind admissions policies by constructing cohorts of applicants with the highest predicted scores under the two corresponding models. We find that test-aware policies yield cohorts with average scores that are about 0.06 population standard deviations higher than under test-blind policies, around two thirds of the gap we observe for policies that only consider academic outcomes.

References

Christopher Avery, Lena Shi, and Preston Magouirk. Test-optional college admissions: ACT and SAT scores, applications, and enrollment changes. Technical report, National Bureau of Economic Research, 2025.

Graeme Blair, Jasper Cooper, Alexander Coppock, Macartan Humphreys, and Luke Sonnet. *estimatr: Fast Estimators for Design-Based Inference*, 2025. URL `https://declaredesign.org/r/estimatr/`. R package version 1.0.4.

Claudia Buchmann, Dennis J Condron, and Vincent J Roscigno. Shadow education, american style: Test preparation, the sat and college enrollment. *Social forces*, 89(2):435–461, 2010.

Elizabeth Cascio, Bruce Sacerdote, Doug Staiger, and Michele Tine. Report from working group on the role of standardized test scores in undergraduate admissions. Technical report, Dartmouth College, 2 2024. URL `https://home.dartmouth.edu/sites/home/files/2024-02/sat-undergrad-admissions.pdf`. Accessed on October 14, 2024.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

Raj Chetty, David J Deming, and John N Friedman. Diversifying society's leaders? the causal effects of admission to highly selective private colleges. Technical report, National Bureau of Economic Research, 2023.

College Entrance Examination Board. College bound seniors, 1974–75. Research Report ED124847, College Entrance Examination Board, New York, NY, 1975. ERIC Document ED124847.

Madison Coots, Soroush Saghafian, David Kent, and Sharad Goel. A framework for considering the value of race and ethnicity in estimating disease risk. *Annals of Internal Medicine*, 178, 2025.

Educational Testing Service. GRE concordance tables: Verbal reasoning and quantitative reasoning. Technical report, Educational Testing Service, Princeton, NJ, 2014. URL `https://web.archive.org/web/20140823203631/ets.org/s/gre/pdf/concordance_information.p`

David F Feldon, Kaylee Litson, Brinleigh Cahoon, Zhang Feng, Andrew Walker, and Colby Tofel-Grehl. The predictive validity of the GRE across graduate outcomes: A meta-analysis of trends over time. *The Journal of Higher Education*, 95(1):120–148, 2024.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

John N Friedman, Bruce Sacerdote, Douglas O Staiger, and Michele Tine. Standardized test scores and academic performance at ivy plus colleges. In *AEA Papers and Proceedings*, volume 115, pages 676–681. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2025.

Nikhil Garg, Hannah Li, and Faidra Monachou. Dropping standardized testing for admissions trades off information and access. *Management Science*, 2026.

Joshua Goodman, Oded Gurantz, and Jonathan Smith. Take two! SAT retaking and college enrollment gaps. *American Economic Journal: Economic Policy*, 12(2):115–158, 2020.

Joshua Grossman, Sabina Tomkins, Lindsay Page, and Sharad Goel. The disparate impacts of college admissions policies on asian american applicants. *Scientific Reports*, 14(1):4449, 2024.

Harvard University. Harvard announces return to required testing, 4 2024. URL `https://news.harvard.edu/gazette/story/2024/04/harvard-announces-return-to-required-te` Accessed on October 14, 2024.

Brian Heseung Kim, Elyse Armstrong, Laurel Eckhouse, Mark Freeman, Rodney Hughes, Trent Kajikawa, and Michelle Sinofsky. Deadline updates, 2023–2024: First-year application trends through February 1. Technical report, Common App, 2 2024. URL `https://www.commonapp.org/files/Common-App-Deadline-Updates-2024.02.14.pdf`. Accessed on October 14, 2024.

Nathan R Kuncel and Sarah A Hezlett. Standardized tests predict graduate students' success. *Science*, 315(5815):1080–1081, 2007.

Nathan R Kuncel, Sarah A Hezlett, and Deniz S Ones. A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological bulletin*, 127(1):162, 2001.

Microsoft. Azure AI Document Intelligence, 2024. URL `https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence`.

National Center for Education Statistics. Integrated Postsecondary Education Data System (IPEDS), 2023, 2024. URL `https://nces.ed.gov/ipeds/use-the-data`. Admissions, graduation rates, and institutional characteristics data from the 2023 collection year, representing the 2022-23 academic year.

OpenAI. OpenAI API. `https://platform.openai.com/docs/guides/embeddings`, 2023. Accessed on October 14, 2024.

OpenAI. GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

QS Quacquarelli Symonds. QS World University Rankings 2025, 2024. URL `https://www.topuniversities.com/world-university-rankings`.

Kelly Ochs Rosinger, Karly Sarita Ford, and Junghee Choi. The role of selective college admissions criteria in interrupting or reproducing racial and economic inequities. *The Journal of Higher Education*, 92(1):31–55, 2021.

Jesse M Rothstein. College performance predictions and the SAT. *Journal of Econometrics*, 121(1-2):297–317, 2004.

Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

Bruce Sacerdote, Douglas O Staiger, and Michele Tine. How test optional policies in college admissions disproportionately harm high achieving applicants from disadvantaged backgrounds. Technical report, National Bureau of Economic Research, 2025.

Scholaro, Inc. Scholaro database, 2025. URL `https://www.scholaro.com/db/`. Accessed September 13, 2025.

Senate-Administration Working Group on Admissions. Final report. Technical report, University of California, San Diego Academic Senate, November 2025. URL `https://senate.ucsd.edu/media/740347/sawg-report-on-admissions-review-docs.pdf`.

SFFA v. Harvard. Students for Fair Admissions, Inc., Petitioner, v. President and Fellows of Harvard College. Students for Fair Admissions, Inc., Petitioner, v. University of North Carolina, et al., 2023. `https://www.supremecourt.gov/opinions/22pdf/20-1199_l6gn.pdf`.

A Additional Notes on Data and Methods

We structure the full details of our data and methods in three parts. Section A.1 describes the covariates we use in our predictive models, including the extraction and featurization pipeline used to structure applicants' transcripts and other admissions materials. In Section A.2, we detail the five academic and holistic outcomes we study. Finally, in Section A.3, we provide additional details on our statistical methods, including model fitting, imputation, inference, and estimation of standard errors.

A.1 Covariates

Admissions officers typically consider admissions factors like standardized test scores and historical GPAs alongside a variety of other potentially relevant information contained in application files. We observe a variety of basic information about candidates, including the admissions cycle in which they applied and whether they were admitted and enrolled; concurrent applications to other programs at our partner institution; and self-reported age, gender, race, military service, and other demographics. We also observe applicants' unstructured transcripts, letters of recommendation, resumes, and application essays, which, however, are not readily amenable to inclusion in predictive models. We develop an LLM-based extraction pipeline to extract structured covariates from these unstructured materials. Below, we detail the featurization process for each category of application material in turn. We extract 396 applicant-level features in total. (Because applicants are required to submit multiple letters and essays and frequently submit several transcripts, aggregation introduces additional features, resulting in a total that exceeds the number of per-material covariates listed below.)

A.1.1 Test scores

Except during the 2020–2021 and 2021–2022 admissions cycles, when testing requirements were suspended because of the COVID-19 pandemic, our partner program has required applicants to submit standardized test scores as part of their admissions materials. A large majority (86%) submit GRE scores, though applicants also have the option of submitting GMAT scores instead. (Due to data issues or other unusual circumstances, around 2% of applicants do not have scores from either test in our data.) To place GMAT and GRE scores on a common scale, we convert GMAT quantitative and verbal scores to equivalent percentile scores for the GRE (Educational Testing Service, 2014). In all predictive models incorporating standardized test scores, we include both GRE quantitative and verbal scores as well as an indicator for whether the covariates represent a converted GMAT score. Some applicants in the earliest cohorts submit pre-2011 GRE scores on the 200–800 point scale, rather than the current 130–170 point scale; we use the official concordance table to convert these scores to the appropriate scale (Educational Testing Service, 2014).

Students who are not native English speakers and did not attend an English-language undergraduate institution are additionally required to submit standardized English proficiency test scores from one of several testing agencies as part of their application. Among international applicants, 38% submit TOEFL scores and 10% submit IELTS scores, representing

the overwhelming majority of English proficiency exam results.

A.1.2 Transcripts

Transcripts from previous educational institutions are a rich source of information about applicants' academic performance. The varied format and grading conventions that transcripts exhibit, however, hinder the extraction of detailed information. To address this challenge, we develop the transcript processing pipeline and feature extraction procedure described below.

Preprocessing. Using OpenAI's GPT-5 series of models (OpenAI, 2025), we develop a four-step extraction pipeline to extract structured information from the transcript scans and photographs applicants submit. First, leveraging the models' multimodality, we orient the pages of each PDF and classify its contents at a high level (transcript, diploma, other), and preprocess the images to increase text contrast and remove watermarks. Second, using Microsoft Document Intelligence (Microsoft, 2024), we transcribe each document using OCR. Third, we prompt a language model with both the transcribed text and images of each relevant page to record course-level information (e.g., term, title, course codes, credits, grade, instructor name, and level) along with term- and degree-specific information (e.g., official institution name and address, matriculation and graduation dates, and majors and minors) in a loosely structured format. This step eliminates extraneous information and improves accuracy in the final step, where we again use a language model to structure the course information according to a fixed JSON schema.

Grade conversion. International applicants make up nearly half of the admissions pool, and many submit transcripts with courses not graded on the 4.0 scale. In the final extraction step, we prompt the language model to convert listed grades to the 4.0 scale using the conversion key included with the transcript, if one is available. If the transcript does not provide an official conversion key, we supply the language model with all available conversion keys from the Scholaro database (Scholaro, Inc., 2025) for the relevant country. We prompt the model to attempt the conversion using background knowledge when no appropriate conversion key is available. For continuous grading scales, we linearly interpolate between entries in conversion tables.

Featurization. From the structured transcript data, we programmatically calculate 336 measures of previous academic performance. These include overall GPAs; GPAs within specific subject areas (e.g., mathematics and statistics, economics) and course levels (e.g., upper- and lower-undergraduate); GPA trends over time; performance in specific courses (e.g., introductory micro- and macroeconomics); indicators for majors, minors, and honors; and other similar measures, including undergraduate GPA, one of our focal covariates. For applicants with multiple transcripts, we calculate credit-weighted GPAs across institutions.

University Rankings. An important subset of our transcript covariates comes from the 2025 QS World University Rankings (QS Quacquarelli Symonds, 2024) and IPEDS survey

data (National Center for Education Statistics, 2024). The QS rankings collect ranking information in addition to other institution characteristics, such as size and level of research activity, for approximately 1,500 universities worldwide. All institutions participating in federal financial aid in the U.S. report a variety of information about their student populations to the Department of Education as part of the IPEDS survey. Schools report information including admission rates, graduation rates, financial information, demographic information, average test scores, and employment outcomes. Many applicants are the only or one of a very small number of applicants to have attended a particular institution in our data. To ameliorate the resulting sparsity in statistical models leveraging previous institutions, we therefore include a large number of covariates from both the IPEDS (44) and QS data (46). For applicants who submit multiple transcripts, we weight IPEDS and QS covariates by the duration of their attendance at each institution.

Applicants, official transcripts, QS, and IPEDS rarely use identical names to refer to the same institution. To match transcripts to the relevant QS or IPEDS entries, we first rank all institutions in the same state (IPEDS) or country (QS) according to embedding similarity using OpenAI's `text-embedding-3-large` (OpenAI, 2023). We validate this ranked list using a language model, which we prompt to select an institution from the top matches that corresponds to the official university title and address information contained in the transcript if there is a matching entry, or reject the match otherwise.

To harmonize our two sources of rankings, we model QS ranking scores as a function of IPEDS covariates for institutions appearing in both datasets. Using the fitted model, we then impute ranks for institutions not included in the QS rankings. Finally, we bin the resulting rankings into quintiles as our measure of institutional rank, placing institutions where neither QS nor IPEDS information is available in the bottom quintile.

A.1.3 Letters of recommendation

Applicants to our partner program must obtain three letters of recommendation from former professors or other instructors, professional supervisors, or other individuals who can speak to an applicant's readiness to attend the master's program. The program requires recommenders to respond to twelve Likert-scale questions about applicants' academic preparation, work ethic, personality traits, and other dimensions of academic potential in addition to the letter itself. Drawing on discussions with partner admissions officers, we develop a rubric to encode aspects of letters' contents that members of the admissions team consider important, including: the relevance of a recommender's relationship to the applicant (e.g., a direct supervisor vs. a personal friend); the strength of their endorsement, overall and with respect to academic potential, leadership, and other pertinent dimensions; the recommender's profession, seniority, and employing organization; their apparent depth of knowledge about the candidate and the length of their relationship; and other similar criteria. We also develop a separate rubric to evaluate whether the letter makes reference to specific applied quantitative skills or other evidence of quantitative preparation. As in the case of transcripts, we use Microsoft Document Intelligence (Microsoft, 2024) to extract each letter's text contents, before prompting language models to (1) extract the structured applicant ratings and (2) evaluate it according to our two rubrics. We extract a total of 26 measures from each letter of recommendation in this way.

A.1.4 Resumes

After extracting each resume's text contents using Microsoft Document Intelligence (Microsoft, 2024), we prompt a language model to standardize an applicant's work history. For each position, we classify the applicant's title and tenure; their employment status (intern, part-time, full-time) and seniority; the sector (for-profit, non-profit, government), size, and prominence of their employer; whether the position involved leadership, public service, or quantitative responsibilities; and other key information. We also prompt the model to classify the applicant's career trajectory according to the promotions, demotions, and lateral career moves contained in their resume. From each applicant's structured work history, we calculate eight summary measures of program-relevant work experience.

A.1.5 Essays

Across the nine admissions cycles comprising our data, our partner program required applicants to submit between two and five essays of around 250 to 500 words, in addition to several optional essays. While the precise prompts vary across application cycles, the core set of topics applicants address remains relatively stable: career goals and reasons for applying to the master's program; descriptions of personal history and background; and demonstrations of grit, resilience, and leadership. Drawing on rubrics used by partner admissions officers, we develop a rubric to evaluate seven aspects of the essays: their (1) grammatical and mechanical correctness, (2) stylistic quality, (3) clarity and organization, (4) adherence to the prompt and formal requirements, (5) voice and originality, (6) public service merit, and (7) evidence of leadership potential. We prompt language models to grade each essay according to our rubric, leveraging concrete evaluation criteria for each dimension, for instance, rating grammatical quality according to the number and severity of errors, to minimize noise in the evaluation process.

A.1.6 Covariate sets

Aside from our focal GPA and GRE score covariates, we collect the remaining features described above into four distinct collections. Specifications with *minimal covariates* contain only focal covariates. Our collection of *baseline covariates* augments the focal covariates with information about an applicant's previous educational institution from IPEDS (National Center for Education Statistics, 2024) or QS (QS Quacquarelli Symonds, 2024), as well as basic administrative information about joint applications to other programs at our partner institution. The collection of *transcript covariates* further augments the baseline covariates with the full set of measures derived from applicant transcripts. Finally, our set of *comprehensive covariates* aims to encompass as much information about applicants as is practicable to include, augmenting the transcript covariates with the full set of features extracted from letters of recommendation, resumes, and essays, along with TOEFL and IELTS scores.

A.2 Outcomes

We study five measures of students' academic and holistic performance, all derived wholly or in part from course grades, describing the construction of these outcomes in detail below. We begin by noting several relevant facts about grading practices at our partner institution. Students at our partner program are graded according to two systems. Letter-graded courses, including all required courses (except during the Spring 2021 term; see below), are graded on the standard 4.0 scale. Certain electives are also graded on a credit-only basis. Some students enroll in courses in other divisions of the institution with different grading systems, which we convert to passing or failing grades according to the standards published by the division to which the course belongs. If a student fails a course and is required to repeat it, we use only the grade obtained on their first attempt.

A.2.1 First-semester quantitative GPA

As described in the main text, our primary outcome of interest is students' performance in core statistics and economics courses, taken during the first semester of the first year. These two courses have had relatively stable curricula and a fixed grading curve over the period we examine. We collapse all grades of B- or below into a single category. Grades below a B- constitute formal course failures, requiring students to repeat the course; in practice, such grades are rare, and B- represents the lowest regularly observed grade.

While the courses are required, students can obtain an exemption if they pass a placement test or can demonstrate that their previous coursework has covered the required curriculum, depending on the course. Across the decade we study, 10% of statistics students and 20% of economics students consequently do not take—and do not have a letter grade in—the corresponding course; in total, 92% of students take at least one of the two courses, though only 76% take both. Students who obtain an exemption to the requirement do not take a course that they can be expected to perform well in. As a result, students who obtain exemptions may have artificially lower GPAs relative to their actual academic performance. For this reason, we analyze two distinct versions of first-semester quantitative GPA. In our primary operationalization, which we use in the main text, we use the average grade obtained in whichever of the quantitative courses a student took in their first semester. As an additional robustness check, we repeat our analyses, dropping observations from students who did not complete both courses.

A.2.2 First-year quantitative GPA

In the second semester, students take an additional required course in both statistics and economics. (Beginning with the 2022 cohort, the second-term required statistics course was divided into two separate half-credit, half-semester courses; we weight our GPA calculations accordingly.) Students may obtain exemptions from these courses as well. During the Spring 2020 semester, courses in our partner program defaulted to credit-only grading following the COVID-19 pandemic, though a minority of students still chose to receive letter grades in their spring-semester core quantitative courses. Overall, 94% of students have an observed grade in at least one core class, though only 62% take the full battery of first- and second-semester

courses. For this reason, we repeat our analyses after dropping observations from students whose first-year quantitative grades are incompletely observed as a robustness check.

A.2.3 Overall GPA

Beyond core quantitative courses, master's students in our partner program take a variety of electives, primarily in their second year, and have the opportunity to take courses offered by other divisions of the institution. As an additional outcome, we compute students' overall GPA across all letter-graded courses taken during their enrollment, weighting courses by credits.

A.2.4 Course passage

Admissions officers may hope to admit students who meet a minimum academic threshold, otherwise prioritizing non-academic desiderata, rather than purely admitting the most academically prepared class. To capture this objective, we analyze whether students achieve at least a B grade in all of their quantitative first-year courses as an additional academic measure. (As noted above, formal failures are rare, with a B- representing the lowest regularly observed grade.) We additionally classify as failing any students who withdraw, receive an incomplete grade, or otherwise do not obtain a B grade or better on their first attempt.

A.2.5 Holistic composite score

Admissions officers at our partner program pursue objectives beyond simply identifying the applicants who are strongest academically. Standardized test scores, which are primarily designed to capture academic aptitude, may be less important for a decision maker weighing competing non-academic goals. To address this possibility, we adapt the official admissions rubric used in our program's admissions process to score applicants along four dimensions: academics, leadership, service, and receptiveness.

To generate these scores, we draw on expert readers' evaluations of applications from the most recent round of admissions. For the academic dimension, we rescale our predictions of applicants' first-year quantitative GPA to match the mean and variance of 1–5 scores assigned by the expert readers. For leadership, service, and receptiveness, we model readers' evaluations using linear models with topically relevant covariates drawn from applicants' letters, resumes, and essays. We apply the fitted models to our historical data to generate 1–5 ratings for each dimension, aggregating them to form a composite score between zero and twenty points using our partner program's weights for each category.

A.3 Statistical methods

An institution implementing a GRE-aware or GRE-blind admissions policy faces two important challenges. First, admissions decisions must be made for new cohorts of applicants using predictive models trained on historical data. Second, academic or other outcomes are typically only observable for past applicants who ultimately enrolled. The goal of our statistical approach, detailed below, is to accurately estimate the consequences of pursuing GRE-aware or GRE-blind admissions under these constraints.

A.3.1 Model fitting and prediction

We fit two classes of predictive models: penalized ridge regression models using the `glmnet` `R` package (Friedman et al., 2010), and gradient boosted decision trees (GBMs) using the `xgboost` `R` package (Chen and Guestrin, 2016). We average predictions from ten independently trained GBMs to reduce noise from the stochastic model fitting algorithm.

After adding indicators for missingness, the number of observations available for model fitting does not greatly exceed the number of features available for generating predictions, a regime in which even linear models are prone to overfitting their data. To address this, we split our admissions dataset by admissions round, using nested cross-validation to tune model parameters and fit predictive models on eight of the nine rounds before generating out-of-sample predictions on each held-out round. We use these predictions to calculate R^2 values, simulate different admissions policies, and estimate other reported quantities. Under the assumption that admissions rounds are exchangeable, this approach mirrors the real-world setting described above, in which historical data must be used to form predictions for new cohorts of applicants.

These predictions can exhibit miscalibration. To correct for this, we compute calibrated predictions using linear or Platt rescaling of the raw model predictions. Because rescaling preserves the relative rank of applicants, it does not affect the simulated admissions policies we consider. The R^2 values we report are likewise invariant.[3] We use calibrated predictions in Figure 3, however, because they are more easily comparable between rounds.

A.3.2 Imputation

The primary statistical challenge in our setting is that grades and other outcomes are only available for applicants who were admitted and attended. Analyses based only on matriculated students can yield skewed conclusions if those applicants differ systematically in unobserved ways from the broader applicant pool. To account for this possibility, we use multiple imputation (Rubin, 1987), modeling outcomes for non-matriculating applicants using the broadest available range of covariates. We also replicate our analyses over two other populations of applicants: an "enrolled" population of students whose outcomes we observe, and therefore need not impute; and an "overlap" subpopulation of applicants who resemble the admitted population.

Imputation models. We train GBMs to predict each outcome using comprehensive covariates supplemented with demographic information. Because multiple imputation requires drawing outcomes from the conditional distribution of outcomes given observables, we model

[3]That is, $\textsc{Cor}(Y, \alpha + \beta \cdot \hat{Y}) = \textsc{Cor}(Y, \hat{Y})$, so our estimates of R^2 are preserved under this rescaling. Similarly, the standard estimator for AUC, which we report for models of course passage, is unaffected by rescaling. However, in general,

$$1 - \frac{\textsc{Var}(Y - [\alpha + \beta \cdot \hat{Y}])}{\textsc{Var}(Y)} \neq 1 - \frac{\textsc{Var}(Y - \hat{Y})}{\textsc{Var}(Y)}.$$

Because the alternate definition of R^2 is impacted by any miscalibration in the predictions $\hat{Y}$, we report both measures in our supplementary analyses.

the full conditional distribution of each outcome rather than the conditional mean only. To account for uncertainty in the hyperparameter tuning and model fitting process and to capture variation across the conditional distribution, we independently tune, fit, and draw from each imputation model 200 times.

- **First-semester quantitative GPA.** We model each student's grade in the required first-semester economics and statistics courses separately as a discrete variable representing the five possible grades in each class (A, A-, B+, B, B- or lower). We then train models predicting students' grades using cross-entropy loss on the full set of students whose grades we observe, tuning model hyperparameters using five-fold cross-validation. To impute missing grades, we draw grades according to the fitted probabilities for both classes and average them to obtain an imputed first-semester quantitative GPA.

- **First-year quantitative and overall GPA.** We model each student's grades across the first-year quantitative requirements as a continuous random variable. We fit models predicting students' GPAs using mean squared error loss. To estimate the conditional residual variance, we then fit a chi-squared (i.e., gamma) regression to the residuals. Because overfitting would tend to optimistically shrink this quantity, we use out-of-sample residuals estimated by splitting observed students into nine folds by admissions round, fitting and tuning a predictive model on eight of the nine folds using five-fold nested cross-validation, and then comparing predicted to actual grades on the held-out round. We then train mean-prediction and variance-prediction models on the entire population of students with observed GPAs, tuning on the same population using five-fold cross-validation. For each individual with a missing GPA, we finally draw a GPA from a normal distribution centered at the predicted mean GPA and with the predicted variance.

- **Course passage.** Since course passage is a binary outcome, we model whether students achieved a sufficiently high grade as a binary variable using log loss. We train our predictive model on the whole set of individuals whose course passage we observe after tuning hyperparameters using five-fold cross-validation on the same set of observations. For individuals with unobserved course passage outcomes, we then draw an imputed outcome according to their modeled probability of passing.

Subpopulations. Estimates based on the entire applicant pool, our *preferred* subpopulation, may suffer from omitted variable bias if applicants who do not enroll differ from enrolled students in unobservable ways. Following existing literature (e.g., Friedman et al., 2025), we therefore repeat our analysis using the subpopulation of *enrolled* students whose outcomes we need not impute. (To ensure comparability across outcomes, we also drop students who are enrolled but missing an outcome.) As a middle ground, we also construct an *overlap* subpopulation of applicants with at least a 10% *ex ante* probability of enrolling, roughly 50% of applicants. We expect imputation to be more accurate for this group because they resemble the pool of enrolled students, whose outcomes we observe. Our model of enrollment, trained in the same way as our other prediction models, achieves an out-of-sample AUC of 78% [CI: 68%–89%], reflecting substantial round-to-round variation in enrollments.

A.3.3 Confidence intervals

Our confidence intervals incorporate three sources of uncertainty: (1) year-to-year variation in the applicant pool and admitted students; (2) randomness in the model tuning and training process; and (3) uncertainty from imputation. For each imputation and model fit, we first calculate the quantity of interest within each admissions round, then estimate its mean and standard error across rounds, treating rounds as exchangeable. (For model performance metrics, this amounts to nine-fold cross-validation with rounds as folds.) We aggregate across imputations and model fits using Rubin's rules (Rubin, 1987) and report t-intervals with eight degrees of freedom to account for the small number of rounds. For Figure 2, which shows linear models fit on the entire population of students with observed first-semester quantitative GPAs, we instead compute t-intervals clustered by round using the `estimatr` `R` package (Blair et al., 2025).

B Mathematical Appendix

Below we prove two theoretical results connecting predictive performance to empirical patterns we observe in different admissions policies. First, in Section B.1, we give the proof of Proposition 1, which connects accuracy measured by a model's R^2 to the academic quality of a class admitted using its predictions. Then, in Section B.2, we introduce a simple model connecting predictive accuracy to the representation of different demographic groups in the admitted class.

B.1 Academic Quality

The proof of Proposition 1 is a straightforward application of well-known identities for conditional normal distributions.

Proof of Prop. 1. Let $\mu \stackrel{\text{def}}{=} \mathbb{E}[Y]$, $\mu_i \stackrel{\text{def}}{=} \mathbb{E}[\hat{Y}_i]$, and $\sigma_i^2 \stackrel{\text{def}}{=} \textsc{Var}(\hat{Y}_i)$. Define $Z \stackrel{\text{def}}{=} (Y-\mu)/\sigma$ and $\hat{Z}_i \stackrel{\text{def}}{=} (\hat{Y}_i - \mu_i)/\sigma_i$ for $i = 0, 1$, so that Z, $\hat{Z}_0$, and $\hat{Z}_1$ are marginally standard normal random variables. Observe that

$$\mathbb{E}[Y \mid \hat{Y}_i > y_i] = \mu + \sigma \cdot \mathbb{E}[Z \mid \hat{Z}_i > \Phi^{-1}(1-p)] = \mu + \sigma \cdot \mathbb{E}[\mathbb{E}[Z \mid \hat{Z}_i] \mid \hat{Z}_i > \Phi^{-1}(1-p)], \quad (2)$$

Here the first equality follows from the definitions of Z and $\hat{Z}_i$ as well as the fact that $\hat{Y}_i > y_i$ if and only if $\hat{Z}_i > \Phi^{-1}(1-p)$. The second equality is a consequence of the law of iterated expectations. Now, observe that $Z \mid \hat{Z}_i \sim \mathcal{N}(\rho_i \cdot \hat{Z}_i, 1-\rho_i^2)$. Consequently,

$$\mathbb{E}[\mathbb{E}[Z \mid \hat{Z}_i] \mid \hat{Z}_i > \Phi^{-1}(1-p)] = \mathbb{E}[\rho_i \cdot \hat{Z}_i \mid \hat{Z}_i > \Phi^{-1}(1-p)] = \rho_i \cdot \frac{\varphi(\Phi^{-1}(1-p))}{p}, \quad (3)$$

where the second equality follows from writing the mean of a truncated normal as the inverse Mills ratio. Since $\Phi^{-1}(1-p) = -\Phi^{-1}(p)$ and $\varphi(-z) = \varphi(z)$, combining Eqs. (2) and (3) and taking the difference for $i = 0, 1$ yields (1). □

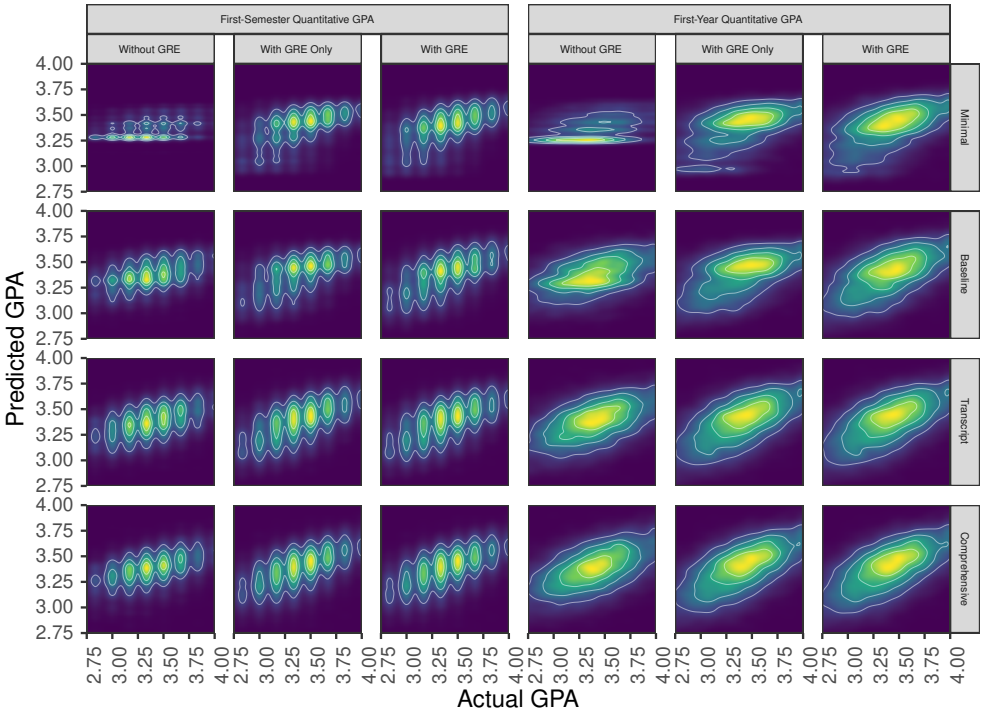

Figure 7: *Joint kernel density estimates of predicted and and actual first-semester and first-year quantitative GPA across focal covariate specifications (GPA-only, GRE-only, and both GPA and GRE scores) and covariate sets. Lighter colors indicate regions of higher density. Each panel shows results from a single imputation.*

We observe that the Mills's ratio inequalities imply that

$$\lim_{p \to 0} \frac{\varphi(\Phi^{-1}(p))}{p \cdot \Phi^{-1}(p)} = -1,$$

i.e., the final factor in (1) diverges like $\Phi^{-1}(p)$ as $p \to 0$. As a result, it is relatively small even for admissions rates much less than even those of the most selective U.S. colleges and universities: For $q = 0.1\%$, $\varphi(\Phi^{-1}(p))/p \approx 3.37$, and for $p = 0.0001\%$, $\varphi(\Phi^{-1}(p))/p \approx 4.95$.

The key assumption in Proposition 1 is that predictions and outcomes are jointly normal. Figure 7 illustrates the degree of joint normality of predictions and outcomes present in our data for both first-semester and first-year quantitative GPA. While first-year GPA is approximately normal across specifications (except when undergraduate GPAs are the only covariate present in the model), first-semester quantitative GPAs exhibit less normality because of the limited number of possible outcomes. Nevertheless, the close fit between empirical and theoretical values shown in Figure 4 suggests that in practice Proposition 1 is reasonably robust to violations of the joint normality assumption.

B.2 Demographic Composition

Like the improvement in the admitted class's academic quality, the impact of increased predictive accuracy on the admitted class's demographic composition can also be understood through a simple statistical heuristic. We model first-semester quantitative GPA or another outcome of interest Y as a mixture of normal distributions:

$$Y \mid G = g \sim \mathcal{N}(\mu_g, \sigma^2), \tag{4}$$

where $G \in \{1, \ldots, m\}$ denotes the group an applicant belongs to, μ_g the average GPA in that group, and $\sigma > 0$ the shared within-group standard deviation of Y. Let $p_g \stackrel{\text{def}}{=} \Pr(G = g)$, where we assume that $p_g > 0$ for all $g = 1, \ldots, m$. Without direct access to Y, admissions decisions are based on predictions $\hat{Y}$ of Y that satisfy

$$Y = \hat{Y} + \varepsilon \tag{5}$$

where the error term $\varepsilon \sim \mathcal{N}(0, \tau^2)$ is independent of $\hat{Y}$ and G. It follows from Cramér's theorem and Eqs. (4) and (5) that

$$\hat{Y} \mid G = g \sim \mathcal{N}(\mu_g, \sigma^2 - \tau^2), \tag{6}$$

i.e., that within each group, $\hat{Y}$ achieves an R^2 of $1 - (\tau/\sigma)^2 = \rho^2$, where $\rho \stackrel{\text{def}}{=} \textsc{Cor}(Y, \hat{Y} \mid G)$ is the within-group correlation between Y and $\hat{Y}$.

If all applicants with a predicted first-semester GPA of at least t are admitted, then it follows from (6) that the proportion of the admitted class belonging to group G satisfies:

$$r_g \stackrel{\text{def}}{=} \Pr(G = g \mid \hat{Y} > t) = \frac{p_g \cdot \bar{\Phi}\left(\frac{t-\mu_g}{\rho\sigma}\right)}{\sum_{h=1}^{m} p_h \cdot \bar{\Phi}\left(\frac{t-\mu_h}{\rho\sigma}\right)}, \tag{7}$$

where $\bar{\Phi}(Z)$ denotes the standard normal CCDF.

The class composition implied by this simple model closely matches what we observe in our data under different admissions scenarios, as shown in Figure 8. Using first-semester quantitative GPA as our outcome of interest, we compare the proportion of admitted applicants we would expect to belong to each quintile of undergraduate institution rank to the value implied by (7) across a range of admissions rates. (Neither the theoretical nor empirical values vary substantially for admissions rates about 25%.) The empirically estimated and theoretically expected values agree well across the range.

Motivated by the empirical accuracy of our simple model, we next turn to understanding how the composition of the class varies as the accuracy of predictions improves. To that end, we consider r_g as a function of the decision threshold t and prediction quality ρ in (7). Because the number of students who can be admitted is usually essentially fixed, rather than a specific admissions threshold t, one can instead consider an admissions probability $q \in (0, 1)$ which, along with ρ, determines the admissions threshold.

Figure 9 illustrates the class composition implied by (7) when we choose thresholds t to fix q while allowing the predictive accuracy ρ to vary. When we compare our empirical

estimates of class composition under models with different predictive accuracies, we again observe reasonably close agreement, with theoretical estimates falling within our estimates' margin of error in most cases, or otherwise deviating by relatively modest amounts. (Estimates using minimal covariates deviate more dramatically, due to more extreme violations of the distributional assumptions, and are not shown; cf. 7. Estimates using transcript and comprehensive covariates are extremely similar.)

Figure 9 exhibits an important empirical phenomenon, also displayed in Figure 6: Using policies based on less accurate models tends to result in class compositions more heavily weighted towards groups with higher average GPAs. To better understand this phenomenon, we begin by proving the following lemma, which characterizes how r_g changes as ρ varies for any fixed q.

Lemma S2. *Fix $q \in (0,1)$. For each $\rho \in (0,1)$, there exists $t_q(\rho)$ satisfying $\Pr(\hat{Y} > t_q(\rho)) = q$, where $\textsc{Cor}(\hat{Y}, Y \mid G) = \rho$. Then as a function of ρ, r_g is differentiable on $(0,1)$, and*

$$\frac{\partial r_g}{\partial \rho} = \frac{p_g}{q\rho} \cdot \varphi(z_g) \cdot \left[z_g - \frac{\sum_{h=1}^m p_h \cdot \varphi(z_h) \cdot z_h}{\sum_{h=1}^m p_h \cdot \varphi(z_h)} \right], \tag{8}$$

where $z_g \stackrel{\text{def}}{=} (t_q(\rho) - \mu_g)/(\rho\sigma)$.

Proof. The proof is a straightforward application of the implicit function theorem. Consider

$$f(\rho, t) \stackrel{\text{def}}{=} \Pr(\hat{Y} > t) = \sum_{g=1}^m p_g \cdot \bar{\Phi}\left(\frac{t - \mu_g}{\rho\sigma} \right).$$

For any fixed $\rho \in (0,1)$, the function $t \mapsto f(\rho, t)$ is continuous and strictly decreasing with

$$\lim_{t \to -\infty} f(\rho, t) = 1, \qquad \text{and} \qquad \lim_{t \to \infty} f(\rho, t) = 0.$$

Therefore, for any $q \in (0,1)$, there exists a unique $t_q(\rho) \in \mathbb{R}$ such that $f(\rho, t_q(\rho)) = q$.

Now, observe that

$$\frac{\partial}{\partial \rho} f(\rho, t) = \sum_{g=1}^m p_g \cdot \varphi\left(\frac{t - \mu_g}{\rho\sigma} \right) \cdot \frac{t - \mu_g}{\rho^2 \sigma}, \qquad \text{and} \quad \frac{\partial}{\partial t} f(\rho, t) = -\sum_{g=1}^m p_g \cdot \varphi\left(\frac{t - \mu_g}{\rho\sigma} \right) \cdot \frac{1}{\rho\sigma}.$$

Since $\partial f / \partial t$ is non-zero, the implicit function theorem implies that the function $\rho \mapsto t_q(\rho)$ is differentiable on $(0,1)$ and

$$\frac{dt_q}{d\rho} = \frac{1}{\rho} \cdot \frac{\sum_{g=1}^m p_g \cdot \varphi\left(\frac{t_q(\rho) - \mu_g}{\rho\sigma} \right) (t_q(\rho) - \mu_g)}{\sum_{g=1}^m p_g \cdot \varphi\left(\frac{t_q(\rho) - \mu_g}{\rho\sigma} \right)}. \tag{9}$$

Because $\Pr(\hat{Y} > t_q(\rho)) = q$, by (7),

$$r_g(\rho) = \frac{p_g}{q} \cdot \bar{\Phi}\left(\frac{t_q(\rho) - \mu_g}{\rho\sigma} \right).$$

Therefore

$$\frac{dr_g}{d\rho} = \frac{p_g}{q}\left[\varphi\left(\frac{t_q(\rho)-\mu_g}{\rho\sigma}\right)\cdot\frac{t_q(\rho)-\mu_g}{\rho^2\sigma} - \varphi\left(\frac{t_q(\rho)-\mu_g}{\rho\sigma}\right)\cdot\frac{1}{\rho\sigma}\cdot\frac{dt_q}{d\rho}\right].$$

Substituting (9) and recalling the definition of z_g gives (8). □

We observe that by (8) the sign of $\partial r_g/\partial\rho$ is entirely determined by

$$z_g - \frac{\sum_h p_h\cdot\varphi(z_h)\cdot z_h}{\sum_h p_h\cdot\varphi(z_h)},$$

i.e., the difference between z_g and a weighted average of $\{z_1,\ldots,z_m\}$ where all the weights are positive. Such a weighted average must lie strictly between the maximum and minimum values, which gives the following corollary.

Corollary S3. *Let $g_{\max}$ and $g_{\min}$ denote the maximum and minimum arguments of μ_g for $g\in\{1,\ldots,m\}$, and suppose that $\mu_{g_{\max}} > \mu_{g_{\min}}$. Then $\partial r_{g_{\max}}/\partial\rho < 0$ and $\partial r_{g_{\min}}/\partial\rho > 0$.*

Proof. Since $\mu_{g_{\max}} \geq \mu_g$ for all $g\in\{1,\ldots,m\}$, $z_{g_{\max}} \leq z_g$ for all $g\in\{1,\ldots,m\}$; similarly, $z_{g_{\min}} \geq z_g$ for all $g\in\{1,\ldots,m\}$. Since $z_{g_{\max}} < z_{g_{\min}}$ and $p_g\cdot\varphi(z_g) > 0$ for all $g\in\{1,\ldots,m\}$, it immediately follows that

$$z_{g_{\max}} < \frac{\sum_{g=1}^m p_g\cdot\varphi(z_g)\cdot z_g}{\sum_{g=1}^m p_g\cdot\varphi(z_g)} < z_{g_{\min}},$$

and so the result follows from (8). □

In our admissions context, Corollary S3 implies that increasing the accuracy of predictions should tend to increase the proportion of admits from the group with the lowest average GPA and decrease the proportion from the group with the highest, exactly as we observe in Figures 6 and 9. Beyond these two groups, the proportion of admits from other groups can increase or decrease, in general. As a simple heuristic, if q is small and the groups are comparable sizes, which is often the case of interest, $t_q(\rho)$ will tend to be larger than μ_g for all or most g. Since $\varphi(z)$ decays to zero rapidly as $|z|\to\infty$, the weights $p_g\cdot\varphi(z)$ will consequently be significantly larger for the highest mean groups, biasing the weighted mean toward the groups with the highest average GPAs. As a result, all but the highest mean groups will typically increase their representation in the admitted class as the accuracy of predictions improves.

Corollary S3 also implies that the most equal class composition results from a perfect predictor of the outcome of interest. We investigate the extent to which predictive uncertainty reduces the representation of students from lower-ranking schools in Figure 10, which compares the proportion of students admitted from each ranking quintile using predictions made with or without GRE scores to the proportion of students who *would* be admitted if admissions officers were able to perfectly forecast future academic performance ("Oracle"). Consistent with our observations above, predictions made using minimal covariates do not follow the expected pattern, due to their non-normality. In the case of comprehensive

covariates, roughly double the number of students from the lowest ranking quintile of undergraduate institutions could be admitted with perfect predictions of future performance, largely replacing students from top-ranked institutions who if admitted would ultimately underperform. In absolute terms, however, the gap is only a few percentage points, and the inclusion or exclusion of test scores does little to close it.

C Robustness Checks

We assess the robustness of our main findings by varying five dimensions of the analysis:

1. **Outcome**: In addition to first-semester quantitative GPA, we also consider policies based on models predicting first-year quantitative GPA, overall GPA, course passage, and our holistic composite score. We also compare first-year and first-semester quantitative GPAs against the alternative variants detailed in Section 2.

2. **Model**: We fit both ridge regression and gradient boosted decision tree models.

3. **Covariate sets**: Beyond comprehensive and baseline covariates, we also consider minimal and transcript covariates.

4. **Subpopulation**: We calculate estimates using the enrolled and overlap subpopulations defined above in addition to the full set of applicants used in our main analysis.

5. **Admission rates**: We simulate both more (10%) and less (30%) selective admissions policies than the 20% admissions rate considered in the main text.

We vary each dimension in turn, holding the other dimensions at their default values (viz., first-semester quantitative GPA, gradient-boosted decision trees, comprehensive covariates, the full applicant pool, and a 20% admissions rate). For each variation, we investigate both academic and demographic outcomes, reproducing Figure 5, Figure 6, and Table 1.

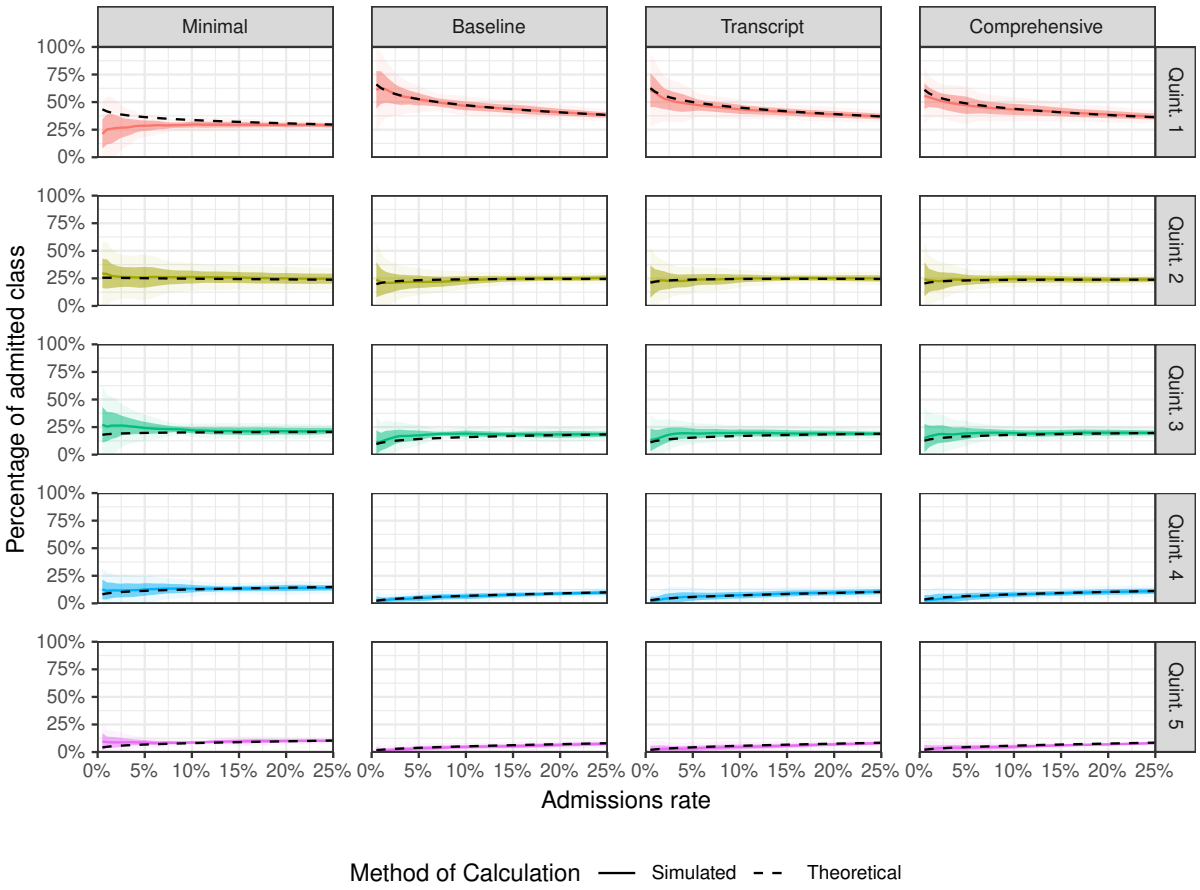

Figure 8: *Class composition implied by* (7)*, shown by dashed lines, versus class composition observed in our policy simulations across a range of admissions rates, shown by solid lines. (For admissions rates greater than 25%, the theoretical and empirical class compositions vary little as a function of the admissions rate and agree closely.) Our admissions policies rank individuals by predicted first-semester quantitative GPA using both undegraduate GPAs and GRE scores. Columns show different covariate sets and rows applicants' undergraduate institution rank, binned by quintile. Dark and light shaded ribbons indicate 68% and 95% pointwise confidence intervals, respectively.*

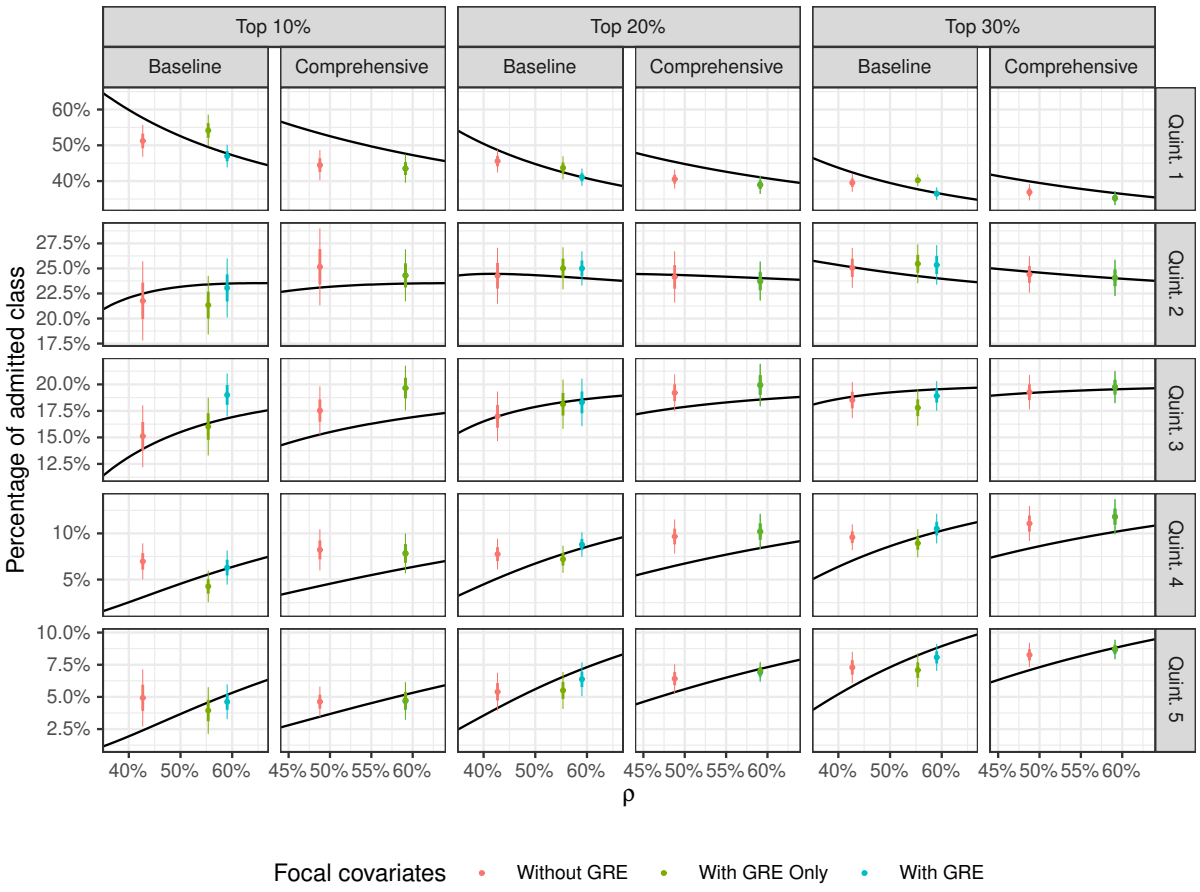

Figure 9: *Class composition implied by* (7) *versus empirical estimates of class compositions for admissions policies based on models achieving different accuracies and with different admissions rates. The colors of the plotted points indicate which focal covariates the predictive model includes: undergraduate GPA only ("Without GRE"), GRE covariates only ("With GRE Only"), or both GPA and GRE covariates ("With GRE"). Admissions rates are indicated by the uppermost facet labels. In facets showing comprehensive covariates, models using GRE covariates only and models using all focal covariates are not separately visible. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

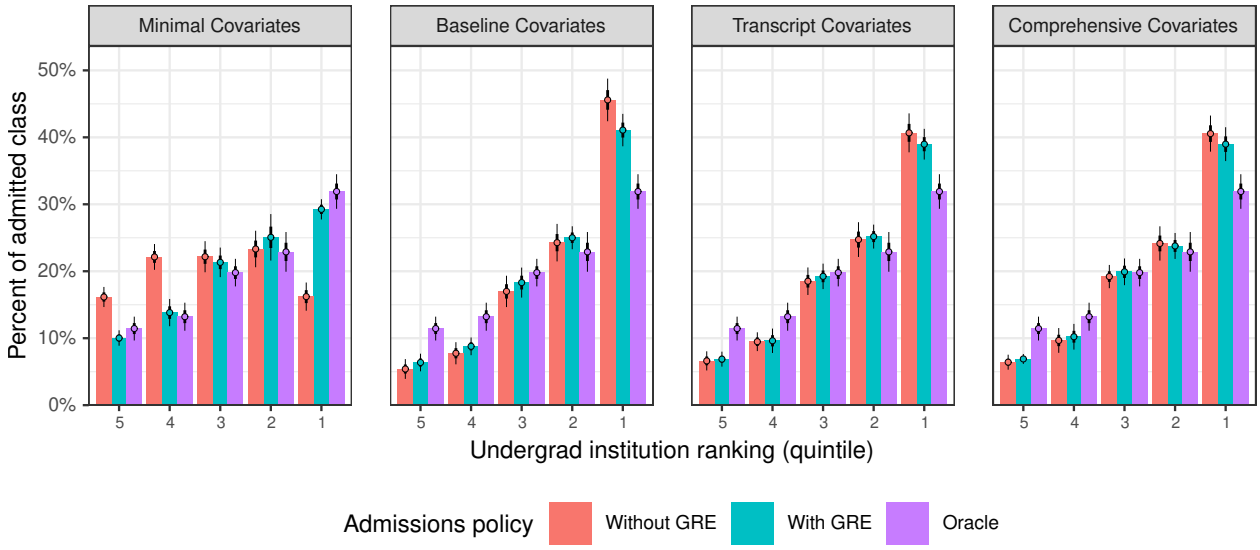

Figure 10: *A comparison of the demographics of students admitted under GRE-aware and GRE-blind admissions policies with an admissions policy using perfect predictions of future grades ("Oracle"). Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

Table 2: Results for First-Semester Quantitative GPA.

		Without GRE	With GRE	Diff.
Adjusted	$R^{2\dagger}$	20%***	30%***	9.7%**
		(13%, 28%)	(22%, 38%)	(4.5%, 15.0%)
	R^2	0.239***	0.350***	0.111***
		(0.201, 0.277)	(0.308, 0.392)	(0.088, 0.135)
	MAE	0.223***	0.210***	-0.013**
		(0.215, 0.231)	(0.200, 0.219)	(-0.021, -0.006)
Unadjusted	$R^{2\dagger}$	24%***	32%***	7.7%***
		(19%, 30%)	(26%, 38%)	(4.7%, 10.6%)
	R^2	0.248***	0.324***	0.075***
		(0.189, 0.308)	(0.262, 0.385)	(0.045, 0.106)
	MAE	0.242***	0.228***	-0.013***
		(0.229, 0.254)	(0.214, 0.242)	(-0.018, -0.009)
	Class GPA	3.62***	3.64***	0.028***
		(3.58, 3.65)	(3.61, 3.67)	(0.017, 0.040)
	Pct. Decisions Differing			9.5%***
				(8.7%, 10.4%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted values are computed over matriculants without imputation.* $R^{2\dagger}$ *indicates* R^2 *values computed as one minus the ratio of the residual and total sums of squares. We report out-of-sample values for all measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

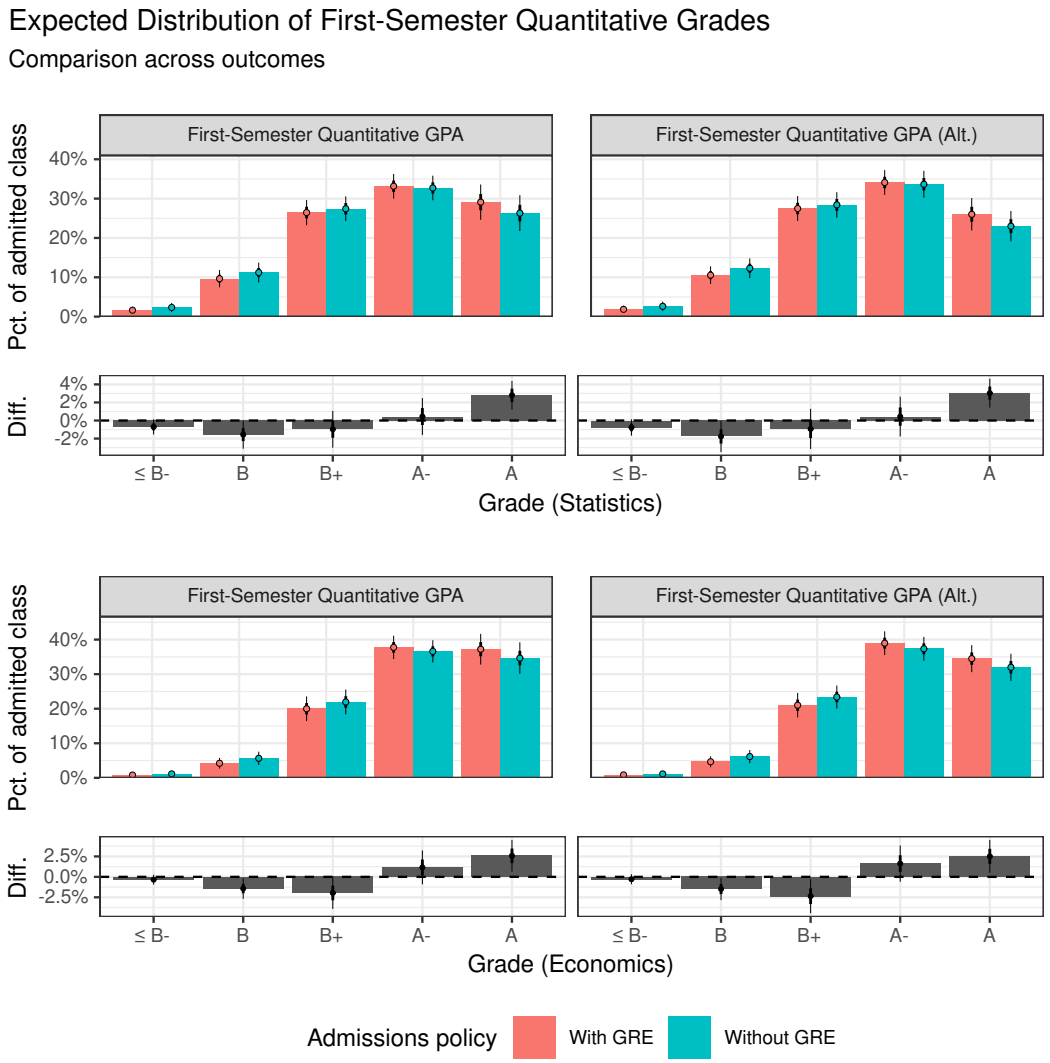

Figure 11: *The distribution of first-semester quantitative GPAs in (simulated) cohorts of students admitted according to their predicted academic performance both with (red) and without (blue) standardized test scores, compared across variant definitions of first-semester quantitative GPA. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

Table 3: Comparison of first-semester quantitative GPA variants.

		Quant. GPA (1st sem.)			Quant. GPA (1st sem., alt.)		
		Without GRE	With GRE	Diff.	Without GRE	With GRE	Diff.
Adjusted	$R^{2\dagger}$	20%***	30%***	9.7%**	17%**	28%***	11.6%***
		(13%, 28%)	(22%, 38%)	(4.5%, 15.0%)	(9%, 25%)	(21%, 36%)	(6.4%, 16.9%)
	R^2	0.239***	0.350***	0.111***	0.208***	0.331***	0.123***
		(0.201, 0.277)	(0.308, 0.392)	(0.088, 0.135)	(0.172, 0.243)	(0.292, 0.370)	(0.101, 0.146)
	MAE	0.223***	0.210***	-0.013**	0.224***	0.209***	-0.015**
		(0.215, 0.231)	(0.200, 0.219)	(-0.021, -0.006)	(0.216, 0.232)	(0.200, 0.218)	(-0.022, -0.008)
Unadjusted	$R^{2\dagger}$	24%***	32%***	7.7%***	20%***	29%***	8.9%**
		(19%, 30%)	(26%, 38%)	(4.7%, 10.6%)	(13%, 27%)	(22%, 36%)	(4.4%, 13.3%)
	R^2	0.248***	0.324***	0.075***	0.209***	0.293***	0.084**
		(0.189, 0.308)	(0.262, 0.385)	(0.045, 0.106)	(0.143, 0.275)	(0.220, 0.366)	(0.041, 0.127)
	MAE	0.242***	0.228***	-0.013***	0.243***	0.229***	-0.014***
		(0.229, 0.254)	(0.214, 0.242)	(-0.018, -0.009)	(0.230, 0.256)	(0.212, 0.245)	(-0.021, -0.008)
	Class GPA	3.62***	3.64***	0.028***	3.59***	3.62***	0.030***
		(3.58, 3.65)	(3.61, 3.67)	(0.017, 0.040)	(3.56, 3.62)	(3.59, 3.65)	(0.019, 0.042)
	Pct. Decisions Differing			9.5%***			10.7%***
				(8.7%, 10.4%)			(9.7%, 11.6%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted values are computed over matriculants without imputation.* $R^{2\dagger}$ *indicates* R^2 *values computed as one minus the ratio of the residual and total sums of squares. We report out-of-sample values for all measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

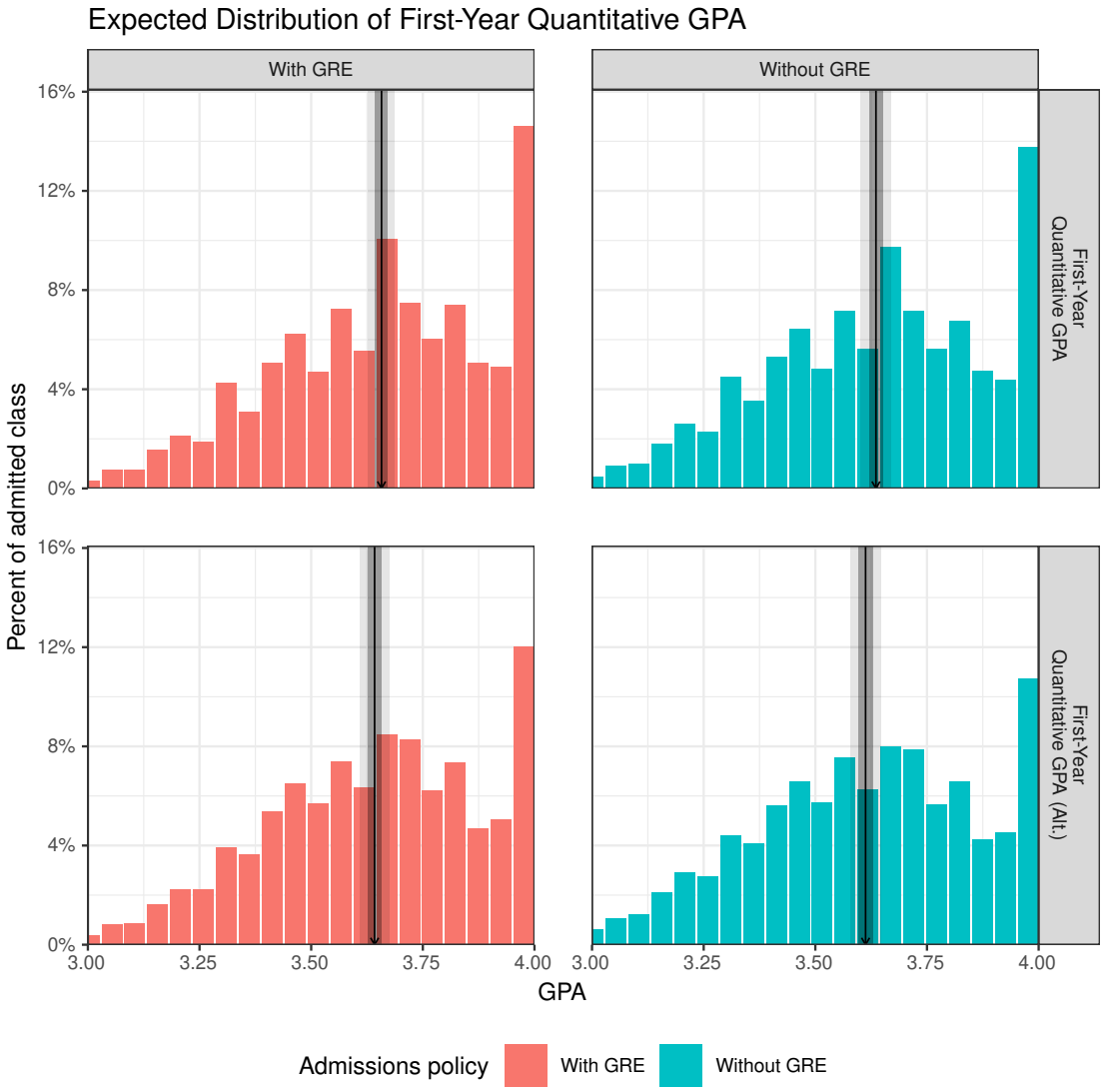

Figure 12: *The distribution of first-year quantitative GPAs in (simulated) cohorts of students admitted according to their predicted academic performance both with (red) and without (blue) standardized test scores, compared across variant definitions of first-year quantitative GPA. The black line shows the admitted class's average GPA under each policy. Dark and light shaded regions and thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

Table 4: Comparison of first-year quantitative GPA variants.

		Quant. GPA (1st year)			Quant. GPA (1st year, alt.)		
		Without GRE	With GRE	Diff.	Without GRE	With GRE	Diff.
Adjusted	$R^{2\dagger}$	30%***	39%***	9.5%***	28%***	38%***	10.5%***
		(27%, 33%)	(36%, 43%)	(7.8%, 11.3%)	(23%, 32%)	(33%, 43%)	(8.6%, 12.3%)
	R^2	0.303***	0.396***	0.093***	0.281***	0.384***	0.103***
		(0.267, 0.339)	(0.361, 0.431)	(0.076, 0.110)	(0.237, 0.326)	(0.337, 0.431)	(0.087, 0.119)
	MAE	0.228***	0.210***	-0.017***	0.227***	0.208***	-0.018***
		(0.220, 0.236)	(0.204, 0.217)	(-0.022, -0.012)	(0.219, 0.235)	(0.201, 0.215)	(-0.023, -0.013)
Unadjusted	$R^{2\dagger}$	28%***	35%***	6.8%***	24%***	32%***	7.9%**
		(24%, 33%)	(30%, 40%)	(3.7%, 9.9%)	(20%, 29%)	(26%, 39%)	(4.2%, 11.7%)
	R^2	0.290***	0.360***	0.070***	0.252***	0.334***	0.082***
		(0.246, 0.333)	(0.309, 0.410)	(0.038, 0.102)	(0.205, 0.300)	(0.266, 0.402)	(0.047, 0.116)
	MAE	0.215***	0.206***	-0.009**	0.218***	0.207***	-0.010**
		(0.206, 0.225)	(0.198, 0.215)	(-0.014, -0.003)	(0.209, 0.226)	(0.197, 0.217)	(-0.017, -0.004)
	Class GPA	3.64***	3.66***	0.022**	3.61***	3.64***	0.030***
		(3.60, 3.67)	(3.63, 3.69)	(0.010, 0.033)	(3.58, 3.65)	(3.61, 3.68)	(0.018, 0.041)
	Pct. Decisions Differing			8.9%***			9.9%***
				(8.0%, 9.8%)			(8.9%, 10.8%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted values are computed over matriculants without imputation.* $R^{2\dagger}$ *indicates* R^2 *values computed as one minus the ratio of the residual and total sums of squares. We report out-of-sample values for all measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

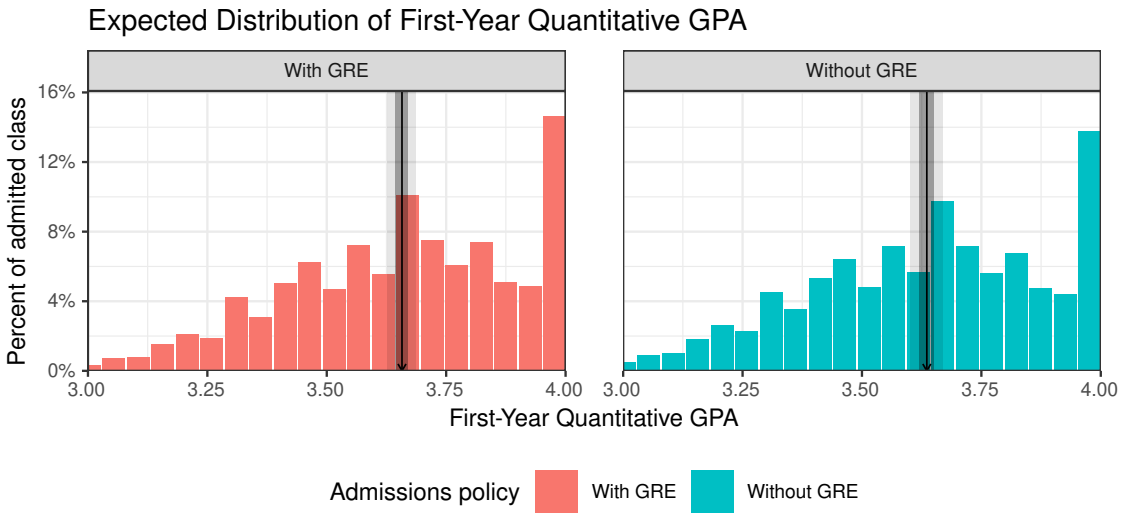

Figure 13: *The distribution of first-year quantitative GPAs in (simulated) cohorts of students admitted according to their predicted academic performance both with (red) and without (blue) standardized test scores. The black line shows the admitted class's average GPA under each policy. Dark and light shaded regions and thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

Table 5: Results for First-Year Quantitative GPA.

		Without GRE	With GRE	Diff.
Adjusted	$R^{2\dagger}$	30%***	39%***	9.5%***
		(27%, 33%)	(36%, 43%)	(7.8%, 11.3%)
	R^2	0.303***	0.396***	0.093***
		(0.267, 0.339)	(0.361, 0.431)	(0.076, 0.110)
	MAE	0.228***	0.210***	-0.017***
		(0.220, 0.236)	(0.204, 0.217)	(-0.022, -0.012)
Unadjusted	$R^{2\dagger}$	28%***	35%***	6.8%***
		(24%, 33%)	(30%, 40%)	(3.7%, 9.9%)
	R^2	0.290***	0.360***	0.070***
		(0.246, 0.333)	(0.309, 0.410)	(0.038, 0.102)
	MAE	0.215***	0.206***	-0.009**
		(0.206, 0.225)	(0.198, 0.215)	(-0.014, -0.003)
	Class GPA	3.64***	3.66***	0.022**
		(3.60, 3.67)	(3.63, 3.69)	(0.010, 0.033)
	Pct. Decisions Differing			8.9%***
				(8.0%, 9.8%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted values are computed over matriculants without imputation.* $R^{2\dagger}$ *indicates* R^2 *values computed as one minus the ratio of the residual and total sums of squares. We report out-of-sample values for all measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

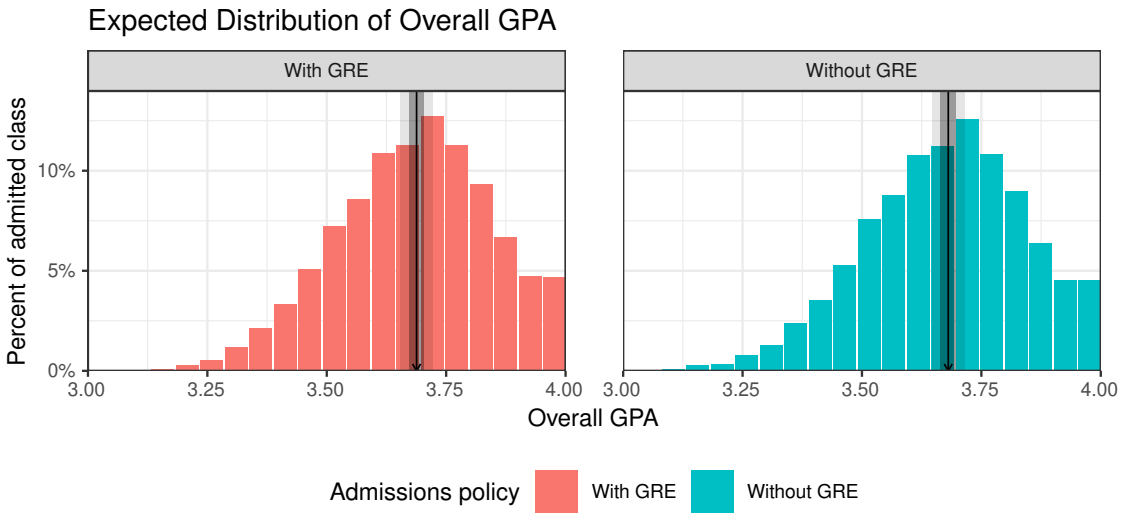

Figure 14: *The distribution of overall GPAs in (simulated) cohorts of students admitted according to their predicted academic performance both with (red) and without (blue) standardized test scores. The black line shows the admitted class's average GPA under each policy. Dark and light shaded regions and thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

Table 6: Results for Overall GPA.

		Without GRE	With GRE	Diff.
Adjusted	$R^{2\dagger}$	23%***	26%***	3.0%***
		(19%, 27%)	(21%, 31%)	(1.6%, 4.3%)
	R^2	0.235***	0.265***	0.030***
		(0.193, 0.277)	(0.218, 0.312)	(0.018, 0.042)
	MAE	0.141***	0.139***	-0.003*
		(0.137, 0.146)	(0.134, 0.143)	(-0.005, -0.001)
Unadjusted	$R^{2\dagger}$	25%***	29%***	3.5%**
		(21%, 30%)	(24%, 33%)	(1.6%, 5.4%)
	R^2	0.261***	0.299***	0.038**
		(0.209, 0.312)	(0.246, 0.351)	(0.017, 0.058)
	MAE	0.135***	0.133***	-0.003*
		(0.129, 0.141)	(0.126, 0.139)	(-0.005, -0.001)
	Class GPA	3.68***	3.69***	0.007*
		(3.65, 3.72)	(3.65, 3.72)	(0.001, 0.013)
	Pct. Decisions Differing			7.0%***
				(6.0%, 8.0%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted values are computed over matriculants without imputation.* $R^{2\dagger}$ *indicates* R^2 *values computed as one minus the ratio of the residual and total sums of squares. We report out-of-sample values for all measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

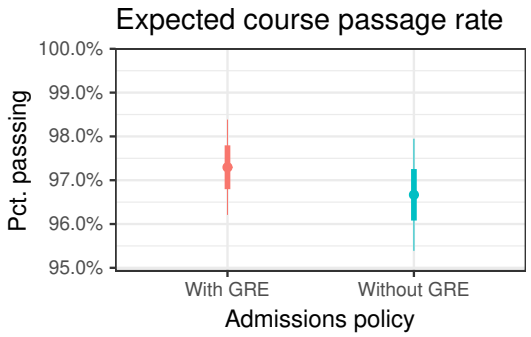

Figure 15: *The expected course passage rate in (simulated) cohorts of students admitted according to their predicted academic performance both with (red) and without (blue) standardized test scores. Thick and thin lines represent 68% and 95% confidence intervals, respectively.*

Table 7: Results for First-Year Course Passage.

		Without GRE	With GRE	Diff.
Adjusted	$R^{2\dagger}$	16%***	22%***	6.7%***
		(12%, 19%)	(18%, 27%)	(4.2%, 9.2%)
	R^2	0.164***	0.226***	0.062***
		(0.128, 0.200)	(0.183, 0.269)	(0.042, 0.082)
	MAE	0.241***	0.237***	-0.004
		(0.218, 0.263)	(0.217, 0.256)	(-0.010, 0.002)
	AUC			
Unadjusted	$R^{2\dagger}$	9%***	13%***	4.3%*
		(5%, 12%)	(8%, 18%)	(0.9%, 7.7%)
	R^2	0.091***	0.136***	0.046*
		(0.061, 0.121)	(0.089, 0.184)	(0.012, 0.079)
	MAE	0.161***	0.153***	-0.007*
		(0.146, 0.175)	(0.138, 0.168)	(-0.014, -0.001)
	Pass Rate	97%***	97%***	0.6%
		(95%, 98%)	(96%, 98%)	(-0.4%, 1.6%)
	Pct. Decisions Differing			8.4%***
				(7.2%, 9.6%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted values are computed over matriculants without imputation.* $R^{2\dagger}$ *indicates* R^2 *values computed as one minus the ratio of the residual and total sums of squares. We report out-of-sample values for all measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

Table 8: Results for Holistic Evaluation.

	Without GRE	With GRE	Diff.
Class GPA	3.58^{***}	3.62^{***}	0.031^{***}
	(3.55, 3.62)	(3.58, 3.65)	(0.019, 0.043)
Class Utility (points)	16.04^{***}	16.22^{***}	0.179^{***}
	(15.81, 16.27)	(15.99, 16.45)	(0.103, 0.255)
Pct. Decisions Differing			$7.8\%^{***}$
			(7.1%, 8.6%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

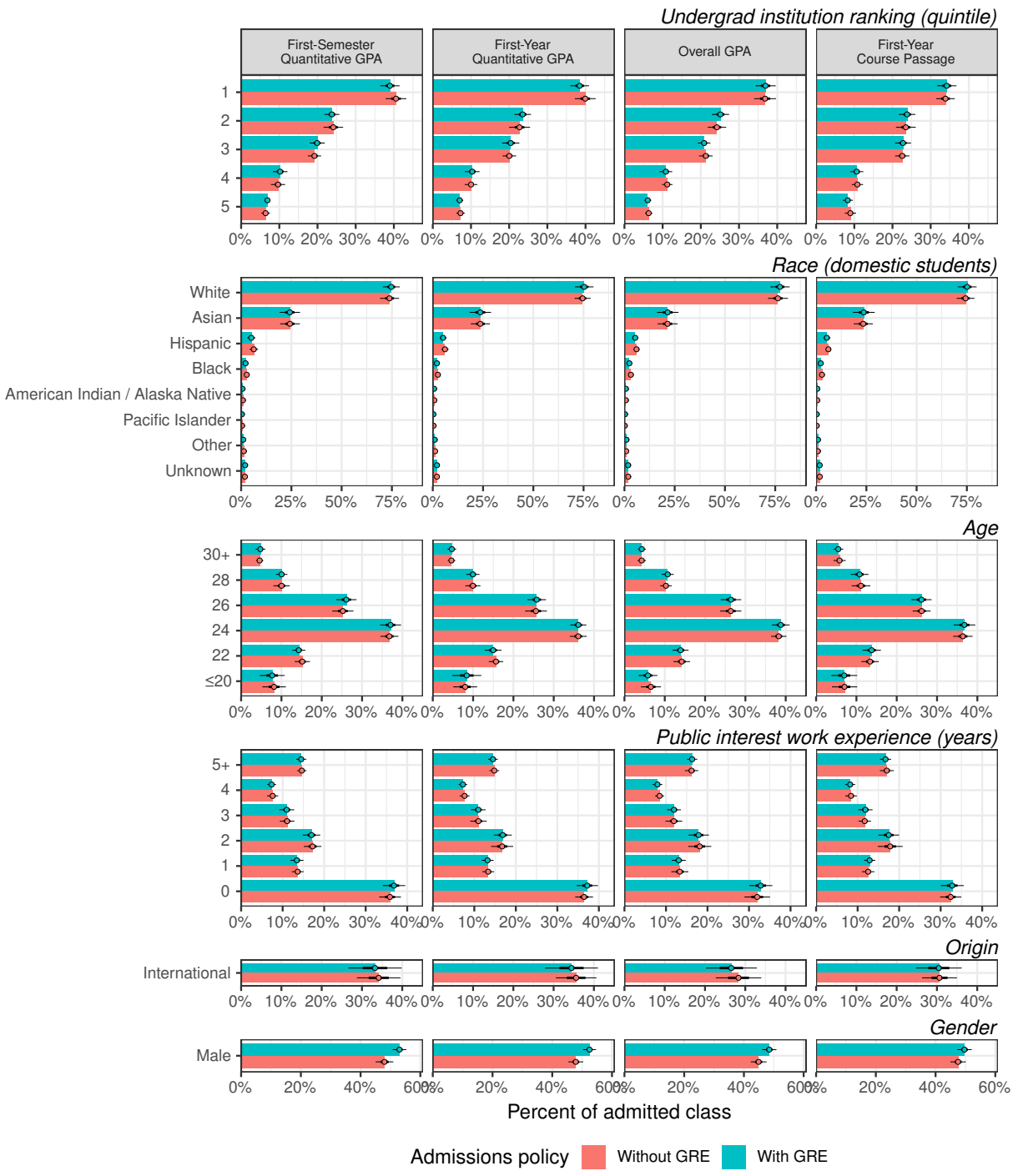

Figure 16: *A comparison of the demographics of students admitted under the GRE-aware and GRE-blind admissions policies using comprehensive covariates, compared across outcomes. The facets show the distribution of admitted students' undergraduate institutional rank, self-reported race (domestic students only), age, work experience, origin, and gender. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

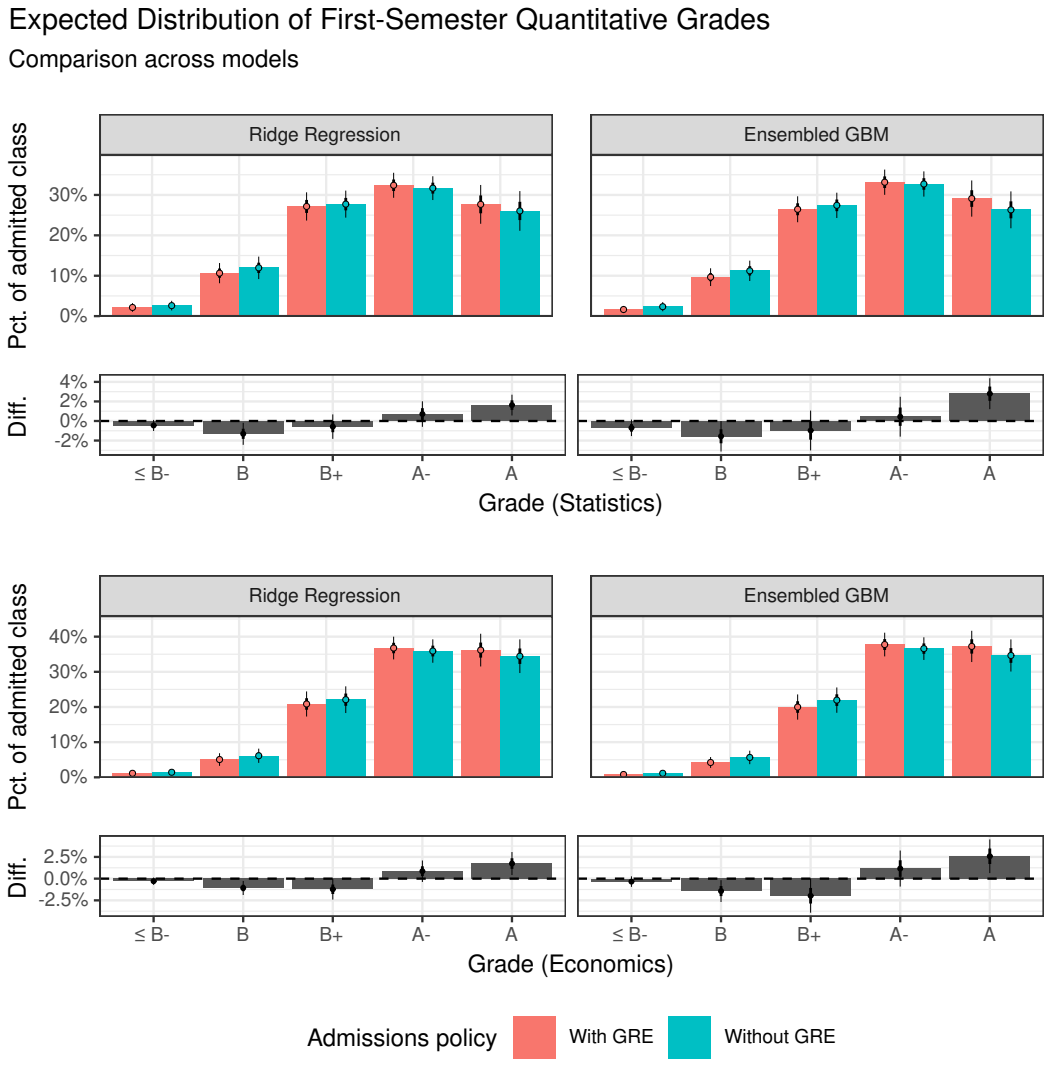

Figure 17: *The distribution of grades in (simulated) cohorts of students admitted according to their predicted academic performance both with (red) and without (blue) standardized test scores, compared across model types. The lower panels of the upper and lower plots show the difference in the proportion of students receiving a given grade under GRE-aware and GRE-blind admissions policies. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

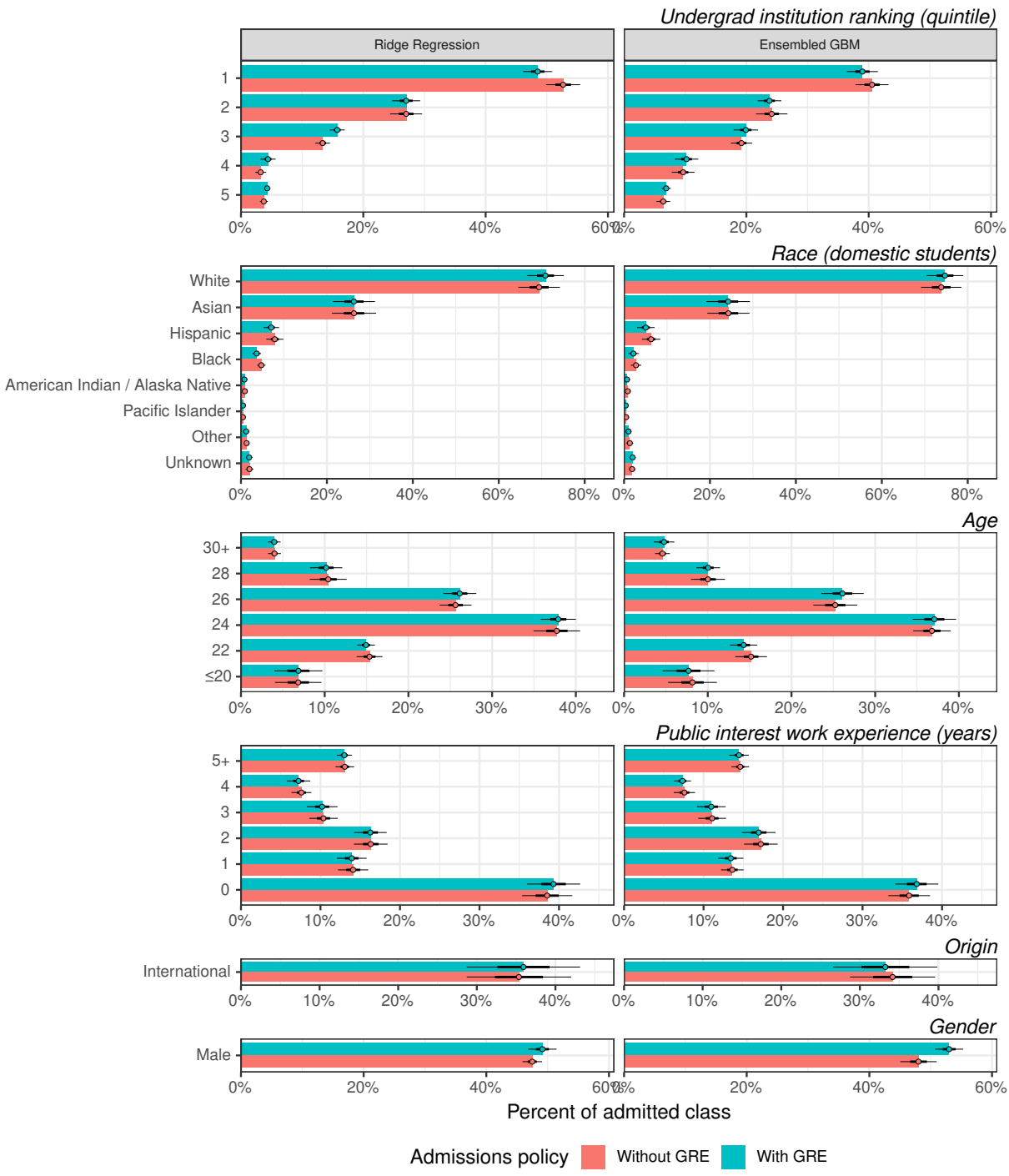

Figure 18: *A comparison of the demographics of students admitted under the GRE-aware and GRE-blind admissions policies using comprehensive covariates, compared across model types. The facets show the distribution of admitted students' undergraduate institutional rank, self-reported race (domestic students only), age, work experience, origin, and gender. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

Table 9: Comparison across models.

		Ridge Regression			Ensembled GBM		
		Without GRE	With GRE	Diff.	Without GRE	With GRE	Diff.
Adjusted	$R^{2\dagger}$	21%***	28%***	6.9%***	20%***	30%***	9.7%**
		(19%, 23%)	(25%, 31%)	(5.9%, 7.9%)	(13%, 28%)	(22%, 38%)	(4.5%, 15.0%)
	R^2	0.231***	0.285***	0.054***	0.239***	0.350***	0.111***
		(0.201, 0.261)	(0.251, 0.319)	(0.047, 0.061)	(0.201, 0.277)	(0.308, 0.392)	(0.088, 0.135)
	MAE	0.230***	0.216***	-0.013***	0.223***	0.210***	-0.013**
		(0.224, 0.236)	(0.211, 0.221)	(-0.016, -0.011)	(0.215, 0.231)	(0.200, 0.219)	(-0.021, -0.006)
Unadjusted	$R^{2\dagger}$	16%***	22%***	5.7%***	24%***	32%***	7.7%***
		(13%, 19%)	(18%, 25%)	(4.9%, 6.5%)	(19%, 30%)	(26%, 38%)	(4.7%, 10.6%)
	R^2	0.210***	0.261***	0.051***	0.248***	0.324***	0.075***
		(0.159, 0.262)	(0.208, 0.315)	(0.039, 0.063)	(0.189, 0.308)	(0.262, 0.385)	(0.045, 0.106)
	MAE	0.259***	0.249***	-0.010***	0.242***	0.228***	-0.013***
		(0.250, 0.268)	(0.239, 0.259)	(-0.011, -0.009)	(0.229, 0.254)	(0.214, 0.242)	(-0.018, -0.009)
	Class GPA	3.61***	3.63***	0.020***	3.62***	3.64***	0.028***
		(3.57, 3.65)	(3.59, 3.66)	(0.011, 0.028)	(3.58, 3.65)	(3.61, 3.67)	(0.017, 0.040)
	Pct. Decisions Differing			3.8%***			9.5%***
				(3.3%, 4.4%)			(8.7%, 10.4%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted values are computed over matriculants without imputation.* $R^{2\dagger}$ *indicates* R^2 *values computed as one minus the ratio of the residual and total sums of squares. We report out-of-sample values for all measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

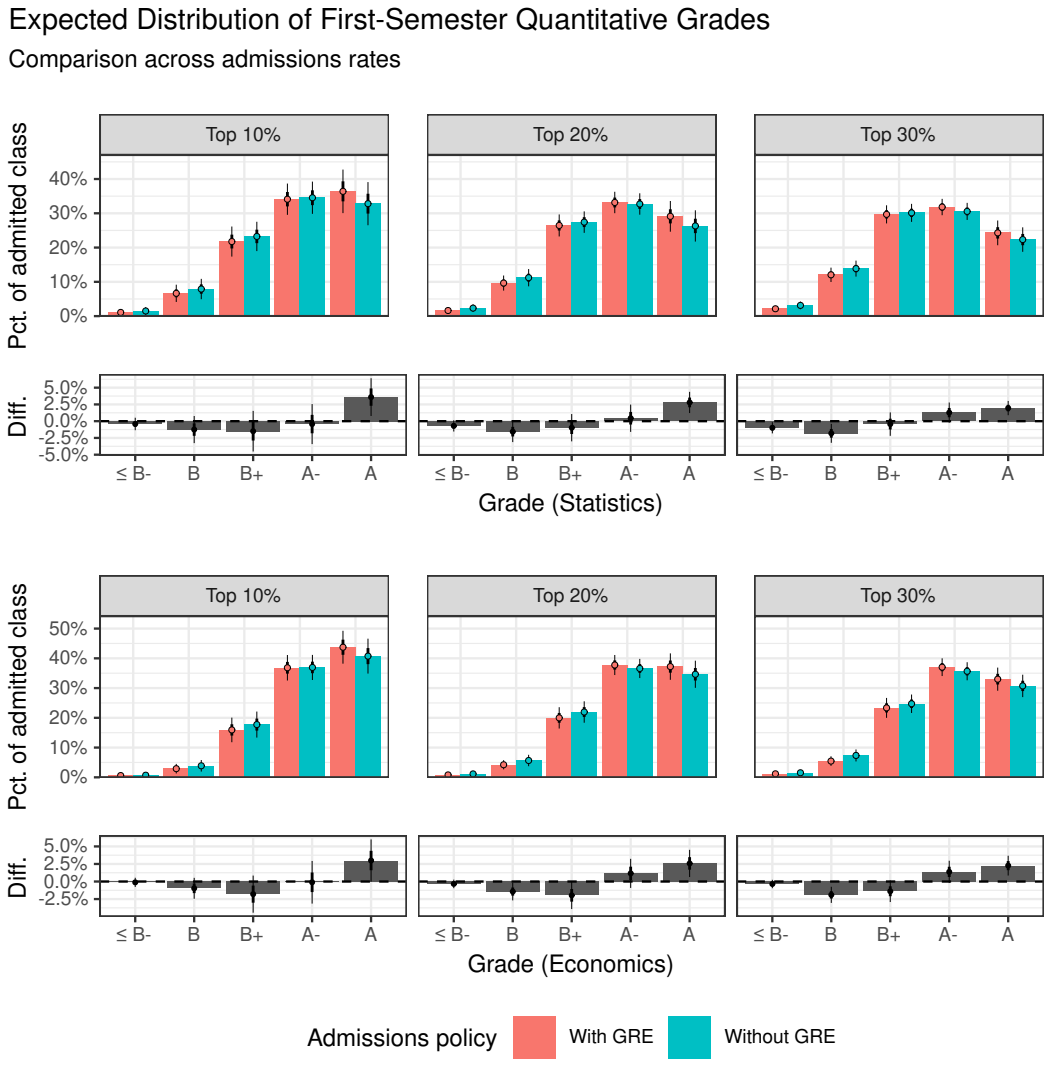

Figure 19: *The distribution of grades in (simulated) cohorts of students admitted according to their predicted academic performance both with (red) and without (blue) standardized test scores, compared across admissions rates. The lower panels of the upper and lower plots show the difference in the proportion of students receiving a given grade under GRE-aware and GRE-blind admissions policies. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

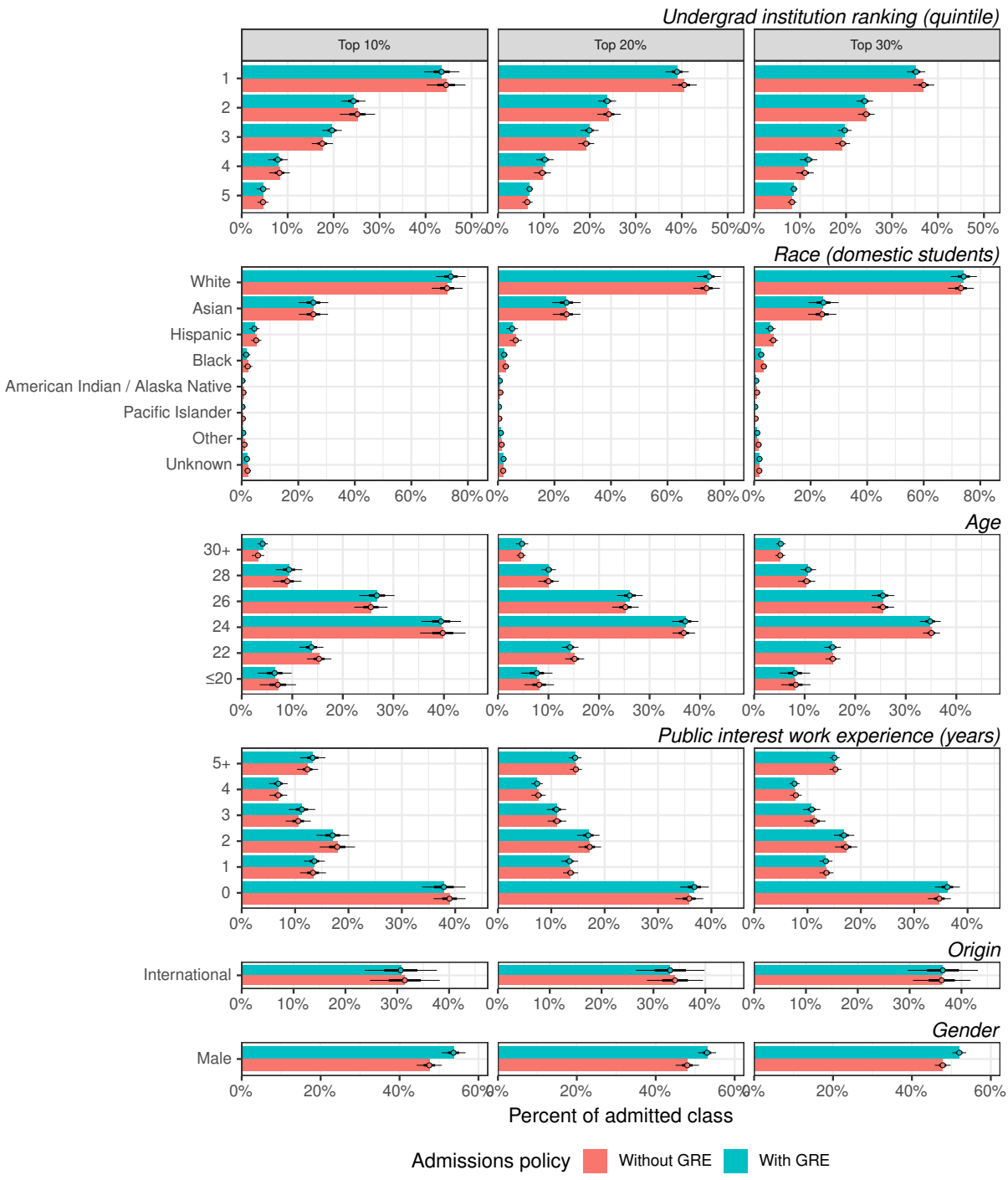

Figure 20: *A comparison of the demographics of students admitted under the GRE-aware and GRE-blind admissions policies using comprehensive covariates, compared across admissions rates. The facets show the distribution of admitted students' undergraduate institutional rank, self-reported race (domestic students only), age, work experience, origin, and gender. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

Table 10: Comparison across admissions rates.

		Top 10%			Top 20%			Top 30%		
		Without GRE	With GRE	Diff.	Without GRE	With GRE	Diff.	Without GRE	With GRE	Diff.
Adjusted	$R^{2\dagger}$	20%***	30%***	9.7%**	20%***	30%***	9.7%**	20%***	30%***	9.7%**
		(13%, 28%)	(22%, 38%)	(4.5%, 15.0%)	(13%, 28%)	(22%, 38%)	(4.5%, 15.0%)	(13%, 28%)	(22%, 38%)	(4.5%, 15.0%)
	R^2	0.239***	0.350***	0.111***	0.239***	0.350***	0.111***	0.239***	0.350***	0.111***
		(0.201, 0.277)	(0.308, 0.392)	(0.088, 0.135)	(0.201, 0.277)	(0.308, 0.392)	(0.088, 0.135)	(0.201, 0.277)	(0.308, 0.392)	(0.088, 0.135)
	MAE	0.223***	0.210***	-0.013**	0.223***	0.210***	-0.013**	0.223***	0.210***	-0.013**
		(0.215, 0.231)	(0.200, 0.219)	(-0.021, -0.006)	(0.215, 0.231)	(0.200, 0.219)	(-0.021, -0.006)	(0.215, 0.231)	(0.200, 0.219)	(-0.021, -0.006)
Unadjusted	$R^{2\dagger}$	24%***	32%***	7.7%***	24%***	32%***	7.7%***	24%***	32%***	7.7%***
		(19%, 30%)	(26%, 38%)	(4.7%, 10.6%)	(19%, 30%)	(26%, 38%)	(4.7%, 10.6%)	(19%, 30%)	(26%, 38%)	(4.7%, 10.6%)
	R^2	0.248***	0.324***	0.075***	0.248***	0.324***	0.075***	0.248***	0.324***	0.075***
		(0.189, 0.308)	(0.262, 0.385)	(0.045, 0.106)	(0.189, 0.308)	(0.262, 0.385)	(0.045, 0.106)	(0.189, 0.308)	(0.262, 0.385)	(0.045, 0.106)
	MAE	0.242***	0.228***	-0.013***	0.242***	0.228***	-0.013***	0.242***	0.228***	-0.013***
		(0.229, 0.254)	(0.214, 0.242)	(-0.018, -0.009)	(0.229, 0.254)	(0.214, 0.242)	(-0.018, -0.009)	(0.229, 0.254)	(0.214, 0.242)	(-0.018, -0.009)
	Class GPA	3.68***	3.70***	0.027**	3.62***	3.64***	0.028***	3.57***	3.60***	0.028***
		(3.64, 3.71)	(3.67, 3.74)	(0.009, 0.044)	(3.58, 3.65)	(3.61, 3.67)	(0.017, 0.040)	(3.54, 3.60)	(3.57, 3.63)	(0.019, 0.037)
	Pct. Decisions Differing			5.6%***			9.5%***			13.1%***
				(4.8%, 6.3%)			(8.7%, 10.4%)			(12.0%, 14.2%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted values are computed over matriculants without imputation.* $R^{2\dagger}$ *indicates* R^2 *values computed as one minus the ratio of the residual and total sums of squares. We report out-of-sample values for all measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

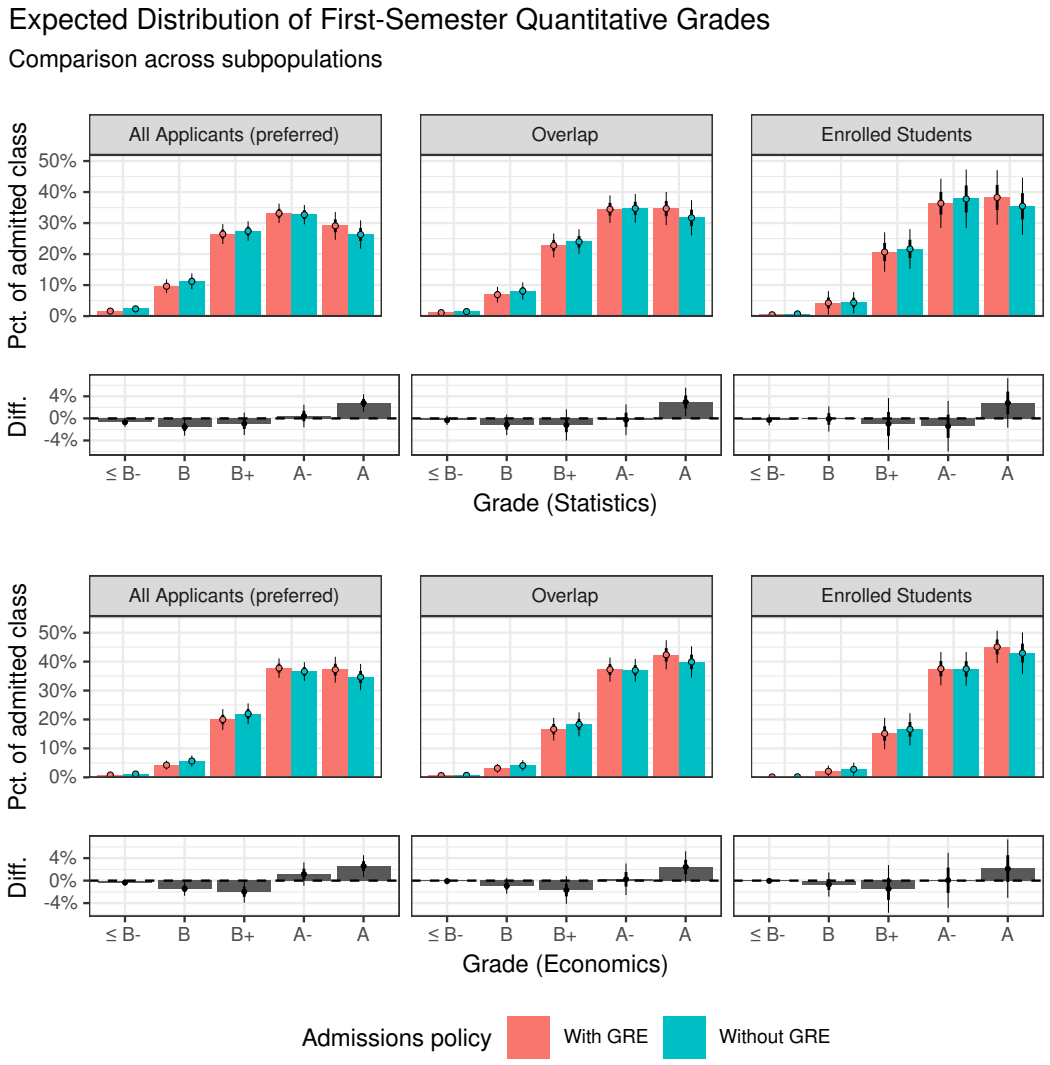

Figure 21: *The distribution of grades in (simulated) cohorts of students admitted according to their predicted academic performance both with (red) and without (blue) standardized test scores, compared across subpopulations. The lower panels of the upper and lower plots show the difference in the proportion of students receiving a given grade under GRE-aware and GRE-blind admissions policies. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

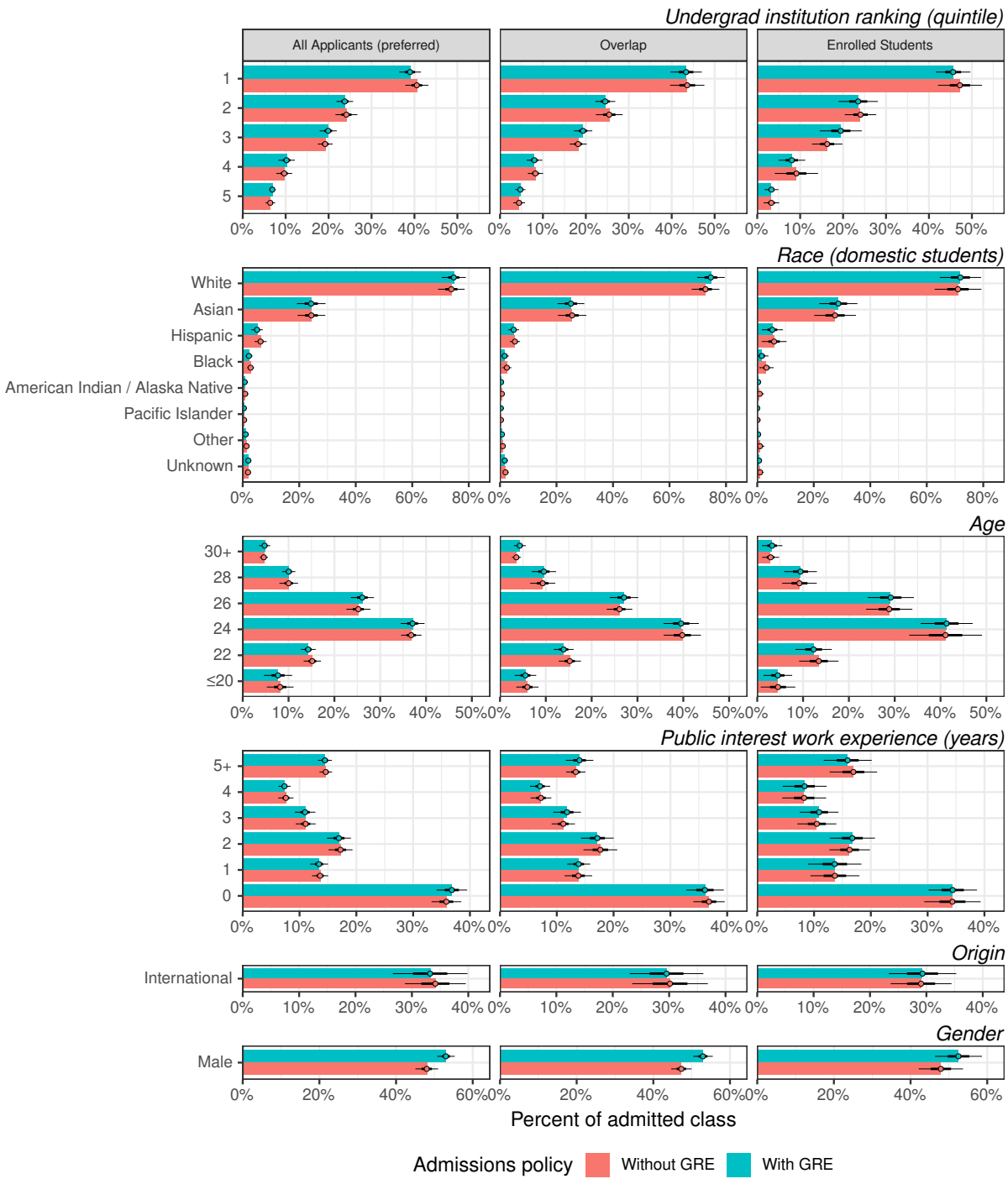

Figure 22: *A comparison of the demographics of students admitted under the GRE-aware and GRE-blind admissions policies using comprehensive covariates, compared across subpopulations. The facets show the distribution of admitted students' undergraduate institutional rank, self-reported race (domestic students only), age, work experience, origin, and gender. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

Table 11: Comparison across subpopulations.

		Preferred			Overlap			Enrolled		
		Without GRE	With GRE	Diff.	Without GRE	With GRE	Diff.	Without GRE	With GRE	Diff.
Adjusted	$R^{2\dagger}$	20%***	30%***	9.7%**	18%***	25%***	7.9%***	25%***	33%***	8.1%***
		(13%, 28%)	(22%, 38%)	(4.5%, 15.0%)	(13%, 22%)	(21%, 30%)	(5.5%, 10.4%)	(20%, 30%)	(27%, 39%)	(5.0%, 11.2%)
	R^2	0.239***	0.350***	0.111***	0.194***	0.269***	0.074***	0.256***	0.336***	0.080***
		(0.201, 0.277)	(0.308, 0.392)	(0.088, 0.135)	(0.162, 0.227)	(0.234, 0.303)	(0.057, 0.092)	(0.200, 0.313)	(0.277, 0.396)	(0.048, 0.112)
	MAE	0.223***	0.210***	-0.013**	0.212***	0.201***	-0.011***	0.240***	0.227***	-0.013***
		(0.215, 0.231)	(0.200, 0.219)	(-0.021, -0.006)	(0.206, 0.218)	(0.195, 0.207)	(-0.015, -0.008)	(0.229, 0.251)	(0.214, 0.239)	(-0.018, -0.009)
Unadjusted	$R^{2\dagger}$	24%***	32%***	7.7%***	21%***	27%***	6.4%**	25%***	33%***	8.1%***
		(19%, 30%)	(26%, 38%)	(4.7%, 10.6%)	(17%, 25%)	(22%, 32%)	(3.2%, 9.7%)	(20%, 30%)	(27%, 39%)	(5.0%, 11.2%)
	R^2	0.248***	0.324***	0.075***	0.214***	0.279***	0.065**	0.256***	0.336***	0.080***
		(0.189, 0.308)	(0.262, 0.385)	(0.045, 0.106)	(0.174, 0.255)	(0.227, 0.331)	(0.033, 0.096)	(0.200, 0.313)	(0.277, 0.396)	(0.048, 0.112)
	MAE	0.242***	0.228***	-0.013***	0.238***	0.228***	-0.010**	0.240***	0.227***	-0.013***
		(0.229, 0.254)	(0.214, 0.242)	(-0.018, -0.009)	(0.227, 0.249)	(0.216, 0.240)	(-0.015, -0.005)	(0.229, 0.251)	(0.214, 0.239)	(-0.018, -0.009)
	Class GPA	3.62***	3.64***	0.028***	3.67***	3.69***	0.022*	3.72***	3.74***	0.014
		(3.58, 3.65)	(3.61, 3.67)	(0.017, 0.040)	(3.63, 3.71)	(3.66, 3.72)	(0.005, 0.039)	(3.68, 3.77)	(3.69, 3.78)	(-0.016, 0.044)
	Pct. Decisions Differing			9.5%***			10.2%***			10.6%***
				(8.7%, 10.4%)			(8.9%, 11.5%)			(8.6%, 12.6%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted values are computed over matriculants without imputation.* $R^{2\dagger}$ *indicates* R^2 *values computed as one minus the ratio of the residual and total sums of squares. We report out-of-sample values for all measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*

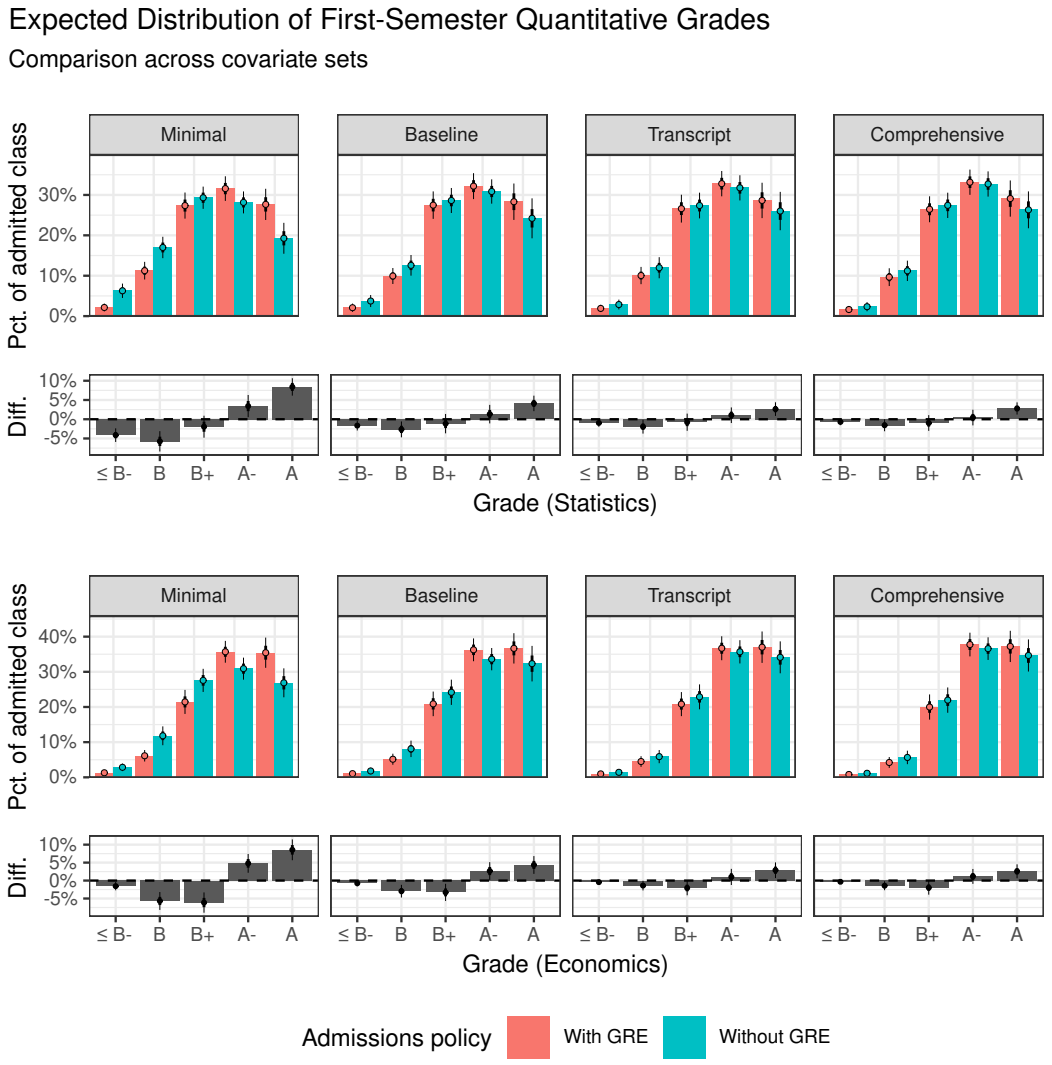

Figure 23: *The distribution of grades in (simulated) cohorts of students admitted according to their predicted academic performance both with (red) and without (blue) standardized test scores, compared across covariate sets. The lower panels of the upper and lower plots show the difference in the proportion of students receiving a given grade under GRE-aware and GRE-blind admissions policies. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

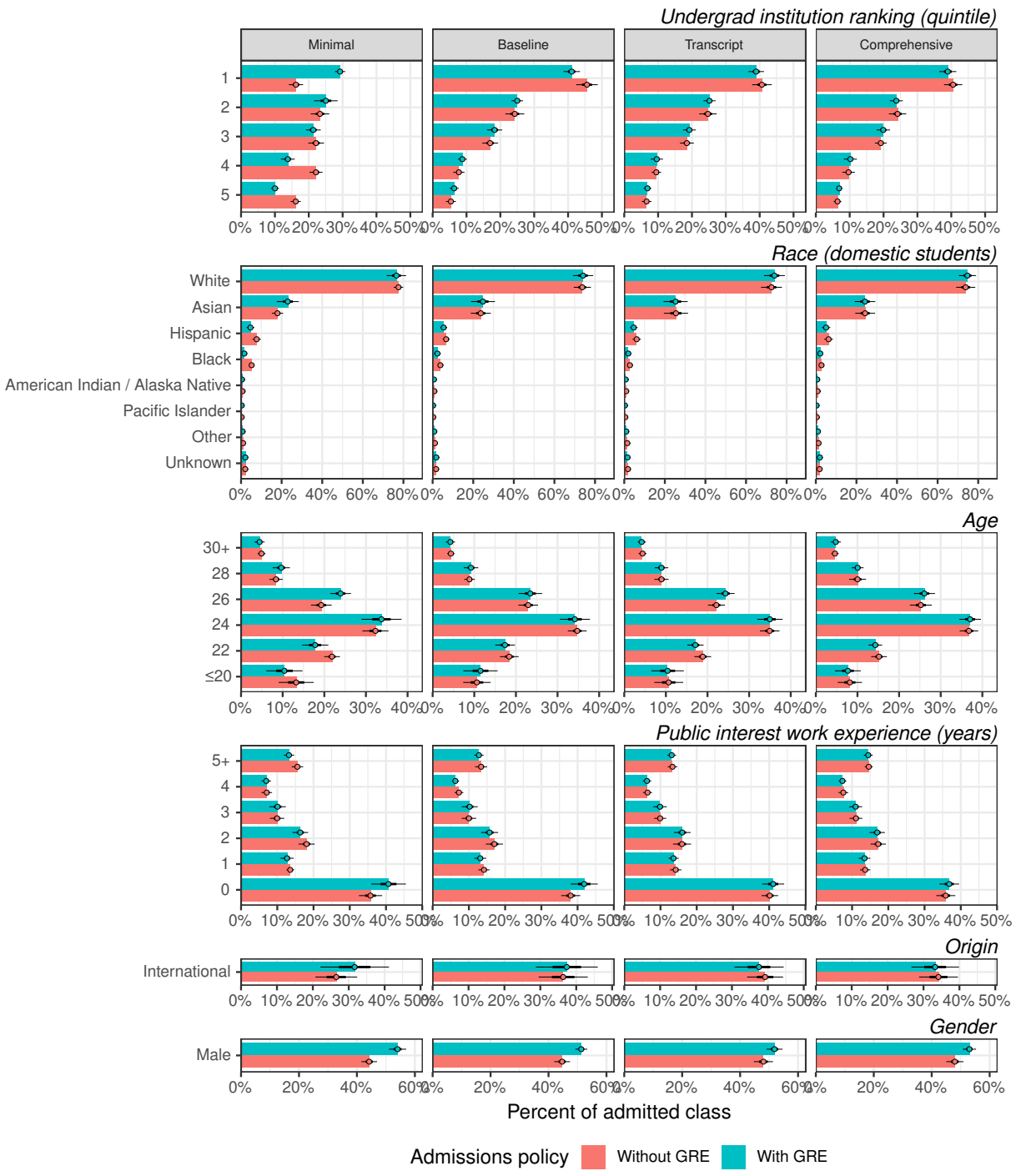

Figure 24: *A comparison of the demographics of students admitted under the GRE-aware and GRE-blind admissions policies using comprehensive covariates, compared across covariate sets. The facets show the distribution of admitted students' undergraduate institutional rank, self-reported race (domestic students only), age, work experience, origin, and gender. Thick and thin lines indicate 68% and 95% confidence intervals, respectively.*

Table 12: Comparison across covariate sets.

	Without GRE	With GRE	Diff.	Without GRE	With GRE	Diff.
	Minimal			**Baseline**		
Range Adj. $R^{2\dagger}$	7%*** (6%, 8%)	27%*** (21%, 33%)	19.8%*** (14.3%, 25.3%)	18%*** (16%, 21%)	33%*** (29%, 38%)	15.0%*** (11.5%, 18.5%)
Range Unadj. $R^{2\dagger}$	5%*** (3%, 6%)	24%*** (19%, 29%)	19.2%*** (14.1%, 24.4%)	15%*** (11%, 19%)	27%*** (22%, 33%)	12.0%*** (7.8%, 16.3%)
Range Adj. R^2	0.075*** (0.060, 0.091)	0.309*** (0.274, 0.344)	0.234*** (0.200, 0.268)	0.183*** (0.160, 0.206)	0.349*** (0.314, 0.384)	0.166*** (0.139, 0.193)
Range Unadj. R^2	0.051*** (0.029, 0.073)	0.245*** (0.192, 0.299)	0.194*** (0.141, 0.246)	0.160*** (0.118, 0.202)	0.280*** (0.226, 0.334)	0.120*** (0.075, 0.165)
MAE	0.253*** (0.245, 0.260)	0.214*** (0.207, 0.221)	-0.038*** (-0.050, -0.027)	0.232*** (0.226, 0.238)	0.205*** (0.198, 0.211)	-0.027*** (-0.035, -0.020)
Class GPA	3.51*** (3.47, 3.54)	3.62*** (3.59, 3.64)	0.108*** (0.083, 0.133)	3.58*** (3.54, 3.62)	3.63*** (3.60, 3.66)	0.051*** (0.031, 0.070)
Pct. Decisions Differing			20.1%*** (19.2%, 20.9%)			13.4%*** (12.4%, 14.3%)
	Transcript			**Comprehensive**		
Range Adj. $R^{2\dagger}$	23%*** (20%, 27%)	33%*** (28%, 38%)	10.0%*** (6.9%, 13.1%)	20%*** (13%, 28%)	30%*** (22%, 38%)	9.7%** (4.5%, 15.0%)
Range Unadj. $R^{2\dagger}$	22%*** (17%, 27%)	30%*** (25%, 36%)	8.5%*** (4.8%, 12.3%)	24%*** (19%, 30%)	32%*** (26%, 38%)	7.7%*** (4.7%, 10.6%)
Range Adj. R^2	0.239*** (0.211, 0.267)	0.360*** (0.323, 0.396)	0.121*** (0.101, 0.142)	0.239*** (0.201, 0.277)	0.350*** (0.308, 0.392)	0.111*** (0.088, 0.135)
Range Unadj. R^2	0.222*** (0.170, 0.274)	0.307*** (0.251, 0.363)	0.085*** (0.047, 0.122)	0.248*** (0.189, 0.308)	0.324*** (0.262, 0.385)	0.075*** (0.045, 0.106)
MAE	0.222*** (0.217, 0.228)	0.204*** (0.197, 0.211)	-0.018*** (-0.025, -0.012)	0.223*** (0.215, 0.231)	0.210*** (0.200, 0.219)	-0.013** (-0.021, -0.006)
Class GPA	3.61*** (3.57, 3.64)	3.64*** (3.61, 3.66)	0.032** (0.017, 0.046)	3.62*** (3.58, 3.65)	3.64*** (3.61, 3.67)	0.028*** (0.017, 0.040)
Pct. Decisions Differing			10.7%*** (9.9%, 11.5%)			9.5%*** (8.7%, 10.4%)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$ *Note: Range adjusted values are computed over the full set of applicants, using imputed values for the outcomes. Range unadjusted values are computed over matriculants without imputation.* $R^{2\dagger}$ *indicates* R^2 *values computed as one minus the ratio of the residual and total sums of squares. We report out-of-sample values for all measures. The proportion of decisions differing is of the entire applicant pool. Parenthetical ranges indicate 95% confidence intervals.*