Ending Affirmative Action Harms Diversity Without Improving Academic Merit

Jinsook Lee*
Cornell University
United States of America
jl3369@cornell.edu

Emma Harvey*
Cornell University
United States of America
evh29@cornell.edu

Joyce Zhou
Cornell University
United States of America
jz549@cornell.edu

Nikhil Garg
Cornell Tech
United States of America
ngarg@cornell.edu

Thorsten Joachims
Cornell University
United States of America
tj@cs.cornell.edu

René F. Kizilcec
Cornell University
United States of America
kizilcec@cornell.edu

Abstract

Each year, selective American colleges sort through tens of thousands of applications to identify a first-year class that displays both **academic merit** and **diversity**. In the 2023-2024 admissions cycle, these colleges faced unprecedented challenges to doing so. First, the **number of applications** has been steadily growing year-over-year. Second, **test-optional policies** that have remained in place since the COVID-19 pandemic limit access to key information that has historically been predictive of academic success. Most recently, longstanding debates over affirmative action culminated in the Supreme Court **banning race-conscious admissions**. Colleges have explored machine learning (ML) models to address the issues of scale and missing test scores, often via ranking algorithms intended to allow human reviewers to focus attention on 'top' applicants. However, the Court's ruling will force changes to these models, which were previously able to consider race as a factor in ranking. There is currently a poor understanding of how these mandated changes will shape applicant ranking algorithms, and, by extension, admitted classes. We seek to address this by **quantifying the impact of different admission policies on the applications prioritized for review**. We show that removing race data from a previously developed applicant ranking algorithm reduces the diversity of the top-ranked pool of applicants without meaningfully increasing the academic merit of that pool. We further *measure the impact of policy change on individuals* by quantifying arbitrariness in applicant rank. We find that any given policy has a high degree of arbitrariness (i.e. at most 9% of applicants are consistently ranked in the top 20%), and that removing race data from the ranking algorithm increases arbitrariness in outcomes for most applicants.

*Both authors contributed equally to this research.

EAAMO '24, October 29–31, 2024, San Luis Potosi, Mexico

ACM ISBN 979-8-4007-1222-7/24/10
https://doi.org/10.1145/3689904.3694706

CCS Concepts

• **Applied computing** → **Law**; • **Social and professional topics** → **User characteristics**; • **Computing methodologies** → **Machine learning**.

Keywords

college admissions, affirmative action, machine learning, ranking system, fairness, arbitrariness

ACM Reference Format:
Jinsook Lee, Emma Harvey, Joyce Zhou, Nikhil Garg, Thorsten Joachims, and René F. Kizilcec. 2024. Ending Affirmative Action Harms Diversity Without Improving Academic Merit. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '24), October 29–31, 2024, San Luis Potosi, Mexico.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3689904.3694706

1 Introduction

At selective American colleges, admissions is a high-stakes decision-making process in which reviewers sort through a large pool of applicants to admit a class that displays both *academic merit* and *diversity* [8, 18]. Many such institutions currently rely on *holistic admissions*, which purport to evaluate the 'whole' student – not only through metrics like test scores and GPA, but also through more subjective factors like background and life experiences [22, 45, 65]. Holistic admissions is a labor-intensive process that requires human review of a complex array of information (grades, activities, essays, letters of recommendation, etc.) for a large volume of applications in a small amount of time.[1] Because the number of applications has been steadily growing year-over-year, this process has become increasingly challenging to scale [48]. A critical question for admissions offices is therefore how to prioritize applications for review in order to most effectively make use of their limited time. In the past, standardized test scores (SAT, ACT) have been used to rank or at least group applicants to organize the review process [36]. However, test-optional policies enacted during the COVID-19 pandemic have limited colleges' access to those standardized metrics, forcing admissions offices to again grapple with the question of how to best organize applicants [39]. Any one piece of information in an application may be too coarse, inconsistently measured, or insufficiently indicative of potential success to impose a ranking on the full applicant

[1] See Appendix B for details of application volumes and timelines at selective American colleges.

pool. Therefore, researchers are increasingly exploring machine learning (ML) approaches to identify complex relationships between historic applications and their corresponding admissions decisions in order to prioritize applicants for review [36, 56, 63, 69, 70].

This raises an important question: what information should these algorithms take into consideration? Prior research efforts have taken inspiration from the ethos of holistic admissions and included all data available in an application, including not only grades, essays, and test scores, but also details of an applicant's background including their race and ethnicity [36]. Until recently, this approach was justifiable under the policy of *affirmative action*, in which applicants belonging to historically marginalized groups were given special consideration in the admissions process. However, in June 2023, affirmative action in college admissions was effectively abolished by the U.S. Supreme Court, which ruled in Students for Fair Admissions v. Harvard that it constituted a form of racial discrimination and was thus unconstitutional [4]. As a result, elements of the admissions process that previously considered race and ethnicity must be updated to exclude those variables. There is currently a poor understanding of how these mandated changes will shape how applicants are prioritized for review, and, by extension, who is admitted.

1.1 Research Questions and Contributions

In this work, we rely on four years of admissions data from a selective American higher education institution to explore how the end of affirmative action is likely to impact admission processes. We address the following research questions:

- **RQ1**: How does a change in admission policy impact a college's *overall class*?
- **RQ2**: How does a change in admission policy impact *individual applicants*?

We cannot observe actual admissions outcomes under different policies. As a proxy, we rely on applicant ranking algorithms, which can be used to determine the order in which applicants are reviewed. We argue that the rankings produced by these algorithms are likely to meaningfully impact admissions. Even if we assume that the order in which an applicant is reviewed does not impact how they are rated – that is, an applicant has roughly the same chance of being considered a 'good' candidate whether they are reviewed first, last, or somewhere in the middle – review order can still impact outcomes. There are constraints on the size of a college's first-year class: there are typically vastly more 'good' candidates than admission slots, meaning that only a subset can be accepted. Based on conversations with the admissions team at the case institution, we model the admissions officers as adding applicants to the pool of accepted students as they review their applications and deem them 'good' candidates. This means that the class may be full before later-reviewed applicants can be added, even if they are deemed equally 'good' by admissions officers. Therefore, we define impact according to *the order in which applicants are prioritized by a ranking algorithm.* Again based on conversations with the case admissions office, we pay special attention to the set of applicants that are designated as part of a 'top' pool of applicants. Finally, incorporating recent scholarship on *model multiplicity* [16] and *arbitrariness* of predictive model decisions [23], we examine the likely impact of policy changes on the outcomes of individual applicants.

Together, this approach allows us to (1) predict the likely impact of race-unaware admissions on the ability of colleges to admit a first-year class that displays both academic merit and diversity; and (2) provide a template for going beyond group fairness assessments of college admissions to understand the relative impact of policy changes on individual applicants. Ultimately, we find that: (1) **race-unaware policies do not meaningfully improve the academic merit of the top-ranked pool even as they significantly decrease diversity**, countering narratives about the costs of affirmative action. At the individual level, we find that (2) **any given policy has a high degree of arbitrariness** (i.e. at most 9% of applicants are consistently ranked in the top 20% by any policy). Further, (3) **arbitrariness in individual outcomes increases under a race-unaware applicant ranking algorithm**.

2 Background and Related Work

In this section, we outline the goals and challenges of modern-day admissions at selective American colleges and how ML has been applied in support of those goals (§2.1). We also provide an overview of how researchers and activists have measured the impact of admissions policies on applicants, highlighting gaps in prior assessment methods that we seek to address with this work (§2.2).[2]

2.1 Selective College Admissions: Goals and Challenges

Admitting applicants with academic merit. Perhaps the most important goal of college admissions is identifying students with academic merit, typically defined as those who are predicted to succeed academically if admitted [7, 9, 68]. As the number of college applications reaches historic highs,[3] this process has become increasingly labor-intensive. Selective colleges must now review tens of thousands of applications in order to fill just a few thousand slots. At the same time, the data available to them to make these decisions is changing. The suspension of standardized tests like the SATs during the COVID-19 pandemic led many colleges to go 'test optional.' Among the approximately one thousand colleges that rely on the Common App, a platform through which millions of college applications in the U.S. are submitted annually, 55% required standardized test scores in 2019, but by 2023, this number plummeted to 4% [34]. Applicants are embracing these test-optional policies: the Common App reports that 76% of applicants submitted test scores in 2019, compared with just 45% in 2023 [34]. As a result, colleges no longer have reliable access to a key piece of nationally standardized information that has been shown to be predictive of student success [21, 25].

Admitting a diverse class. Another goal of college admissions is diversity. For most of American history, access to higher education was largely restricted to those who were white, male, and Protestant. Slavery, coupled with anti-literacy legislation in the South and

[2]We note at the outset that we focus here on *selective, American* colleges and universities, as these are the institutions whose admissions processes are most likely to be impacted by the Supreme Court's recent ban on affirmative action. Colleges outside of the U.S. are not subject to the ruling; non-selective colleges by definition admit the majority of their applicants and as such typically do not consider race in admissions [13, 58].

[3]According to the National Center for Education Statistics, college applications increased by 36% between 2014 and 2022, from 9.6 to 13.1 million [48].

laws barring Black students from public schools in many Northern states, restricted access to education for most Black Americans before the Civil War [50, 51]. Following the abolition of slavery, this discrimination persisted in the form of legally codified segregation that provided Black Americans with lower-quality education. Racial minorities, religious minorities, and women who did apply for higher education were often rejected due to outright bans or quotas restricting their admission; those who were admitted faced segregation and other forms of discrimination [52]. In the 1960s, spurred by nationwide civil rights protests, selective colleges began adopting affirmative action policies to admit Black and other minority students who, as a direct result of this systemic discrimination, did not have comparable grades and test scores to their white peers [66]. Colleges argued that affirmative action was beneficial not only to historically disadvantaged students, but to the institution as a whole due to the "educational benefit that flows from student body diversity" [3]. Although affirmative action has its roots in racial justice movements, many have argued that socioeconomic diversity should be an important consideration as well, pointing out that even selective colleges that have increased enrollment among students of color admit disproportionately few students from low-income families [21, 38]. This view became even more salient in 2023, after decades of legal challenges limiting affirmative action [1, 2] culminated in the Supreme Court ruling that race-conscious admissions "violate the Equal Protection Clause of the 14th Amendment" [4], effectively banning the use of race data in college admissions. We refer to this as the *SFFA*[4] *policy change* throughout.

Machine learning in admissions processes. In the face of these challenges, colleges are increasingly turning to ML to aid their admissions processes [43]. One common, and potentially fraught, use case is to rank applicants using ML (typically by generating scores corresponding to applicants' predicted chance of admission) in order to speed up or scale human review [60, 64, 70]. For example, GRADE, a tool used for graduate admissions at the University of Texas at Austin, was developed because "the number of applications [had] become too large to manage with a traditional review process" [70]. At UT Austin, reviewers were asked to 'validate' applicants to whom GRADE had given very high or low scores and focus most of their time and energy on reviewing applicants about whom GRADE was unsure [70]. GRADE was used for years before it was abandoned in 2020 due to widespread concerns that it was reinforcing historical biases in admissions [19].

However, the use of non-ML based applicant ranking and selection approaches can also be controversial. Even deciding whether to rely on standardized test scores, for example, entails a complex tradeoff [26]: although research has shown that test scores display racial and socioeconomic gaps that may not be reflective of merit [57], there is also evidence that those same scores improve the ability of colleges to identify qualified under-represented applicants [5, 47], and that they may encode less bias than other, more subjective, application materials [24]. In this context, multiple researchers have explored the possibility of using *fairness-aware* ML to improve diversity in admissions. Alvero et al. [10], for example, found that even simple natural language processing (NLP) models were able to distinguish between college application essays written by students of different income levels and genders. Lee et al. [37] subsequently quantified the impact of using this information in an admissions decision support algorithm, finding that essay data helped improve gender diversity, but did not have a significant impact on racial diversity. In a similar study, Lee et al. [36] took a fairness-aware approach to build an applicant ranking algorithm to replace standardized test scores, explicitly considering demographics like race along with other holistic variables in order to increase an institution's ability to identify a diverse set of students with high academic merit. We contribute to this prior work by taking a fairness-aware approach to explore how applicant ranking algorithms based on a broad set of features are likely to change under the SFFA policy change.

Finally, we note that there is a large and growing body of work on domain-agnostic fair ranking algorithms [15, 20, 54, 61, 62, 72–74]. However, the applicant ranking algorithms of which we are aware do not tend to incorporate these approaches [36, 70]; moreover, to the extent that fair ranking mechanisms require access to demographic data, they will not be feasible in college admissions in the future due to the SFFA policy change. As a result, we do not explore fair ranking algorithms in this work, which is intended not to *recommend* approaches for building applicant ranking algorithms, but rather to *predict* how the SFFA policy change will impact already-existing processes. We leave the development of such approaches, that are compatible with the legal environment, for future work.

2.2 Measuring the Impact of Admission Processes and Policies on College Applicants

Impact on groups of applicants. Given the importance of education and the fraught nature of admissions, many researchers have sought to measure the impact of admissions processes on applicants. These assessments often focus on *demographic fairness*, which is typically measured by comparing admission rates across demographic groups like race, gender, and socioeconomic status. In two notable and recent studies, both Grossman et al. [28] and Chetty et al. [21] conducted large-scale analyses on application data to quantify admissions disparities across demographic groups and identify features driving those disparities. Grossman et al. [28] focused on race, finding that Asian students were significantly less likely to be admitted to selective American colleges than white students with comparable test scores, grades, and extracurriculars, and that this disparity was partially (but not entirely) driven by legacy admissions and geographic considerations. Chetty et al. [21] focused on socioeconomic status, finding that students from families with incomes in the top 1% of the U.S. are more likely to be admitted to Ivy-Plus colleges, and that this disparity is mostly driven by factors like recruited athlete status, legacy status, and 'non-academic' (e.g. extracurricular) ratings. Both papers also explore admissions policies that could alleviate these disparities by reconsidering how achievement indicators and sociodemographic attributes are considered in the process [21, 28].

Other researchers have taken more qualitative approaches to examine fairness as well as *transparency* and *trust* in admissions processes. For example, through field observation and a series of interviews at the University of Oxford, Zimdars [75] found evidence that unconscious bias by reviewers led to disproportionately high

[4]After Students for Fair Admission Inc., (SFFA), who brought the suit leading to the Supreme Court ruling.

admission rates for applicants who were white, male, and members of the professional class. Marian [42] crowdsourced data on school assignments in New York City to identify the factors that impacted students' outcomes in order to increase transparency in the process. Similarly, Robertson et al. [59] explored school assignment algorithms in San Francisco through the lens of value-sensitive design in order to understand why changes intended to increase diversity had in practice exacerbated school segregation.

Our contribution: impact on applicants as individuals. Missing from prior assessments is an examination of *arbitrariness* in college admissions. Prior work on *model multiplicity* has shown that predictive tasks can be satisfied by multiple models that are equally accurate (e.g., at identifying applicants with academic merit) but differ in terms of their individual predictions (e.g., whether a specific applicant is ranked at the top) [16, 71]. This means that any number of seemingly minor decisions made by a college admissions office – down to something as simple as how to sample training data – could impact an individual applicant's outcome even if it does not majorly affect the ability of the college to identify students with academic merit. In fact, Cooper et al. [23] show that training otherwise identical models on bootstrapped samples of the same dataset can result in predictions for individuals that are essentially arbitrary: half the time, individuals are predicted to belong to one binary class, and half the time to the other. Similar, seemingly arbitrary, choices, such as how to pre-process variables, have also been shown to cause major changes in model outputs [31]. There is evidence that there are more applicants with high academic merit in a typical pool than there are open seats in a college's first-year class. In particular, a large proportion of applicants with perfect or near-perfect test scores are rejected by selective colleges [28]; additionally selective colleges offer positions on their waitlists to hundreds if not thousands of "[s]tudents who met admission requirements but whose final admission was contingent on space availability" [32].[5] It is likely that if arbitrariness resulting from minor modeling choices impacts how an applicant is prioritized for review by an ML model, then it could also have a downstream impact on their final admissions outcome.[6] We propose that in order to fully understand the impact of admission processes and policies on college applicants, researchers must consider not only fairness and diversity in group outcomes, but also fairness and arbitrariness in individual outcomes.

3 Data and Methods

We seek to quantify and contextualize the impact that the SFFA policy change will likely have on which applicants are prioritized for review (i.e., ranked in the top category by an ML algorithm). In this section, we describe the data available to us and provide a brief overview of the first-year undergraduate admissions process at our case institution (§3.1); we also describe our baseline ranking models (§3.2) and simulated policy changes (§3.3). Finally, we outline our approach to measuring and contextualizing expected changes in the predictive power and diversity of overall rankings (§3.4) and individual applicant outcomes (§3.5) due to the SFFA policy change.

[5]See Appendix B for details of waitlist sizes for selective American colleges.

[6]This follows from the assumption we describe in §1.1: that admissions officers add applicants to the pool of accepted students as they review their applications and deem them 'good' candidates.

3.1 Background and Ranking Algorithm

Case institution. Our case institution is a highly selective, engineering-focused American university. Through the Common App, the case institution collects first-year applicants' standardized test scores, high school grades and coursework, extracurricular activities, family and demographic background (including race and ethnicity, citizenship, and parental education), essays, and letters of recommendation. The case institution has been test-optional since the 2020-2021 admissions cycle. The data we use for this analysis spans all Regular Decision applicants from the 2019-2020 admissions cycle to the 2022-2023 admissions cycle (four years of applications, 59,833 in total). During the 2019-2020 admissions cycle (before the test-optional policy), 92% of applicants submitted either SAT or ACT scores;[7] in the years since, an average of 67% of applicants have submitted either SAT or ACT scores. At the case institution, each application is reviewed twice: first by a seasonal reader, and then by a member of the staff of the admissions office. Applications were historically prioritized for review based on standardized test scores; following the establishment of the test-optional policy, the admissions office implemented ML-generated scores instead.

Preprocessing. To understand how the SFFA policy change is likely to impact *already existing* applicant ranking algorithms, we followed similar data preprocessing steps to those already implemented by the case institution, described in more detail in Appendix C. We included all data that was common across all years, except for personally identifiable information (name, date of birth, contact information, etc.) and data that is not entered directly into the Common App form but is instead provided as a file upload (transcript, letters of recommendation, etc.). We split our data into train and test sets based on year, using the 2019-2020, 2020-2021, and 2021-2022 admissions cycles as training data and the 2022-2023 cycle as test data. This mimics the real-world scenario of training a model on all available historical data, and also allows us to use results for a complete applicant pool as test data. Summary statistics describing our data are presented in Table 1.

Modeling. Finally, we used a Gradient-Boosted Decision Tree to predict applicants' probability of admission.[8] We then segmented applicants into deciles based on that predicted probability: Decile 1 contains the 10% with the lowest predicted probability of admission and Decile 10 contains the 10% with the highest. We then further segment these deciles so that Deciles 9 and 10 (the 20% of applicants with the highest predicted probability of admission) are grouped together as the 'top' pool: this is the set of applicants that the admissions office wants to review first. This follows the approach outlined in Lee et al. [36] and also aligns with how admissions offices are likely to use applicant ranking algorithms in practice:

[7]We assume that the remaining 8% of students submitted their scores late, meaning we do not have access to that data.

[8]We note that the selection of the target variable is a non-trivial decision [53]. We chose to classify applicants who were 'admitted' or 'conditionally admitted' as our positive cases, but provide an analysis of the robustness of our approach with alternative target variable selection in Appendix F, including students who were waitlisted alongside those admitted or conditionally admitted.

Table 1: Summary statistics for our training and test data. The training set includes three years of data from 2019-2020, 2020-2021, and 2021-2022 admissions cycles, and the test set includes the 2022-2023 cycle.

Sample	# applicants	% accepted	% waitlist	% URM	% female	% FG	% LI
Train	44,293	5.7	14.4	17.3	31.1	15.9	25.7
Test	15,540	5.0	16.1	16.3	32.1	19.5	32.3

as a way of prioritizing applications for review, and not as a way of directly admitting applicants.[9]

3.2 Defining Baselines

To measure the impact of the SFFA policy change, we must first establish a baseline. We do this in two ways. First, we train an *ML baseline* model that uses every variable available from the processed Common App data (as described above) to predict an applicant's likelihood of being accepted. Following Lee et al. [36], we believe that this model is a reasonable representation of how college admissions offices might incorporate ML into their processes. We also define a *naive baseline* that ranks applicants based first on their highest level of prior math instruction and then on their standardized test scores.[10]

3.3 Simulating Policy Changes

To analyze the impact of the SFFA policy change and contextualize its impact compared to other hypothetical policy changes, we omitted various variables of interest from the ML baseline. We modeled three different policy changes:

- **No race**: We removed 12 features describing applicants' race, ethnicity, and URM status. These represent the variables that must be excluded from the admissions process due to the SFFA policy change.
- **No major**: We removed 1 feature describing applicants' intended major. This model contextualizes the impact of excluding race by allowing us to compare it to the impact of excluding another important[11] feature that groups applicants but does not directly indicate membership in a historically disadvantaged group.
- **No uncontrollable features**: We removed 29 features representing *uncontrollable* elements of an application (compared to *controllable* elements like major or test scores that an applicant can choose or change). Uncontrollable features include all race-related features, sex, socioeconomic status, citizenship, family education, and type of school attended. This model contextualizes the impact of excluding race from applicant ranking algorithms by allowing us to compare it to the impact of excluding *all* uncontrollable features.

[9]In order to ensure that our results are not brittle to our specific choice of cutoff for assigning the top pool, we verify via a robustness assessment that our findings are consistent across decile cutoffs. The results of that analysis are shown in Appendix E.
[10]We describe the naive baseline in more detail in Appendix D.
[11]Intended major is theoretically important because a smaller percentage of applicants indicating a popular intended major, such as Computer Science, can be admitted given institutional constraints on major size.

3.4 Group Impact: Measuring Changes in Academic Merit and Diversity of the Top-Ranked Applicant Pool

Measuring academic merit of the top pool. An important consideration for any applicant ranking algorithm is whether it can successfully identify applicants with high academic merit who should be prioritized for review by the admissions office. While the 'academic merit' of an applicant is not directly measurable [33], we define two proxies. First, we rely on labels provided by the case institution. We say that the academic merit of a top pool increases as the proportion of applicants who were actually admitted or waitlisted increases. We consider applicants who were not only admitted but also waitlisted because those applicants represent students with high academic merit who would be admitted if there was space for them [32].[12] Second, to isolate a measure of academic merit not dependent on historical decisions, we also use the average percentile ranking of test scores submitted by applicants in the top pool. To measure whether policy changes result in statistically significant differences in the ability of ranking algorithms to identify applicants who were actually admitted or waitlisted, we conduct a binomial test comparing the share of those applicants identified in the top pool by the ML baseline model to the share identified by all other models. Similarly, to measure whether policy changes result in statistically significant differences in the ability of ranking algorithms to identify applicants with high test scores, we used the Mann-Whitney U test to compare across the ML baseline and all other models. For statistical significance, we apply the Benjamini-Hochberg procedure with a false discovery rate of 0.05.[13]

Measuring diversity of the top pool. We define the diversity of our top-ranked pool according to several factors. First, we look at the breakdown of applicants according to their self-identified race and ethnicity (based on the categories available in the Common App). We also consider the share of applicants who are under-represented minorities (URM).[14] Finally, we consider socioeconomic factors, including the share of applicants who identify as being the first in their family to attend college (first-geneneration or FG), and the share of

[12]In fact, in our case institution, the pool of applicants who were admitted *or* waitlisted have equally high test scores as only the pool of applicants who were admitted (Fig. 3). Further, the demographic breakdown of the full pool of applicants is extremely similar to the demographic breakdown of the pool of applicants who were admitted *or* waitlisted, while the URM share is higher among only the pool of applicants who were admitted (Fig. 1). This implies that using an applicant's inclusion in the pool of admitted *or* waitlisted students is likely to be a good indicator of the admissions office's evaluation of that applicant's merit, potentially independent of the demographic considerations that historically could factor into admissions decisions.
[13]We report adjusted p-values throughout.
[14]We define URM in the same way as the Common App, which classifies "Black or African American, Latinx, American Indian or Alaska Native, or Native Hawaiian or Other Pacific Islander" applicants as URM applicants [34].

applicants with low family incomes (low-income or LI).[15] We used binomial tests as described above to determine whether any policy changes result in statistically significant changes to the diversity of the top-ranked pool of applicants as compared to the ML baseline.

It is important to note that the above is a narrow definition of diversity. URM status – and for that matter, racial categorization as defined by the Common App – is controversial and cannot fully represent an applicant's racial identity [11]; further, an applicant's contribution to a diverse campus cannot be reduced to their race/ethnicity and socioeconomic status. However, we choose to focus on these factors as we believe that URM and FGLI status are salient in the context of the SFFA policy change: opponents of the change worry that it will result in a decline in URM enrollment in selective American colleges specifically, and some have suggested mitigating this impact through an increased focus on socioeconomic diversity (i.e. FGLI status) in admissions instead [30, 44].

3.5 Individual Impact: Measuring Changes in Applicant Outcomes

Measuring arbitrariness across models. For an individual applicant, arbitrariness across different ranking algorithms can be measured based on how consistently the applicant is placed in the top pool, or not placed in the top pool. To quantify arbitrariness, we adopted the metric of *self-consistency*, defined by Cooper et al. [23] as "the probability that two models produced by the same learning process on different n-sized training datasets agree on their predictions for the same test instance" [23]. For a given applicant, self-consistency is defined as:

$$\text{SC} = 1 - \frac{2M_0M_1}{M(M-1)} \tag{1}$$

where M is the total number of models we examine, M_1 is the number of models in which an applicant is placed in the top pool and M_0 is the number of models in which an applicant is *not* placed in the top pool. While self-consistency as defined by Cooper et al. [23] measures the consistency of decisions regardless of what those decisions are, we further distinguish between consistently being placed in the top pool and being not placed in that pool throughout.

Predicting how arbitrariness will change as a result of the SFFA policy change. Finally, we consider the fact that arbitrariness is not a fixed quantity: it can increase or decrease across ranking policies (as an intuitive example, the naive baseline has zero arbitrariness across repeated applications; a random ranking policy would be completely arbitrary). We therefore explore how individual arbitrariness resulting from random modeling choices changes if the ML baseline ranking algorithm is minimally modified to comply with the SFFA policy change (i.e. the 'no race' model). We do this by conducting Wilcoxon signed-rank tests that compare the overall arbitrariness and arbitrariness of specific demographic groups between the ML baseline and 'no race' models. As above, we apply the Benjamini-Hochberg procedure with a false discovery rate of 0.05 to account for multiple comparisons.

[15]We do not have direct access to applicants' family incomes status, so we use whether an applicant received an application fee waiver as a proxy; see: https://appsupport.commonapp.org/applicantsupport/s/article/What-do-I-need-to-know-about-the-Common-App-fee-waiver

4 Results

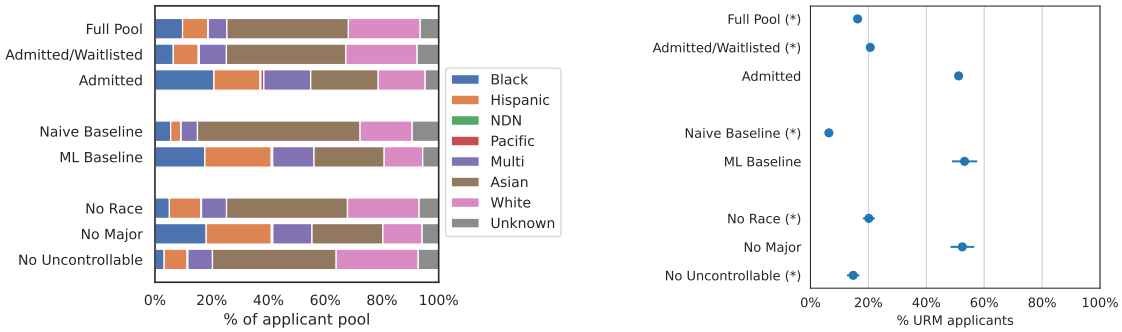
4.1 Group Impact: Academic Merit and Diversity

Compliance with the SFFA policy change significantly reduces the diversity of top-ranked applicants. The ML baseline model, which uses all available Common App data to predict past admissions decisions, represents a reasonable assumption of how admissions offices might previously have implemented applicant ranking algorithms and selects a top-ranked pool that is 53% URM, as shown in Fig. 1. This over-represents URM applicants compared to the full applicant pool (16% URM) and the admitted/waitlisted group (21% URM), but is close to the actual share of URM applicants in the admitted group (51% URM). When we remove data on applicant race and ethnicity from the ML baseline, the URM share in the top-ranked pool drops to 20%—a 62% reduction, which is statistically significant ($p < 0.001$). If we additionally exclude data on other uncontrollable factors like gender and socioeconomic status, the URM share falls even further to 15% ($p < 0.001$). By contrast, excluding major preference from the applicant ranking algorithm results in a URM share of 52%, which is not statistically significantly different from the ML baseline ($p = 0.48$).

Similar trends hold for socioeconomic diversity metrics, as shown in Fig. 2. Excluding data related to race reduces the share of LI applicants in the top pool by a practically and statistically significant amount as compared to the ML baseline (from 31% to 26%, $p < 0.001$), and statistically significantly reduces the share of FG applicants (from 27% to 26%, $p = 0.0496$). Excluding all uncontrollable features further exacerbates the reduction in LI (to 16%, $p < 0.001$) and FG (to 11%, $p < 0.001$) applicants in the top pool. Excluding applicants' intended major has mixed effects on socioeconomic diversity: the share of LI applicants increases to 32%, but this is not statistically significant (p = 0.16); the share of FG applicants increases statistically significantly to 29% ($p = 0.02$).

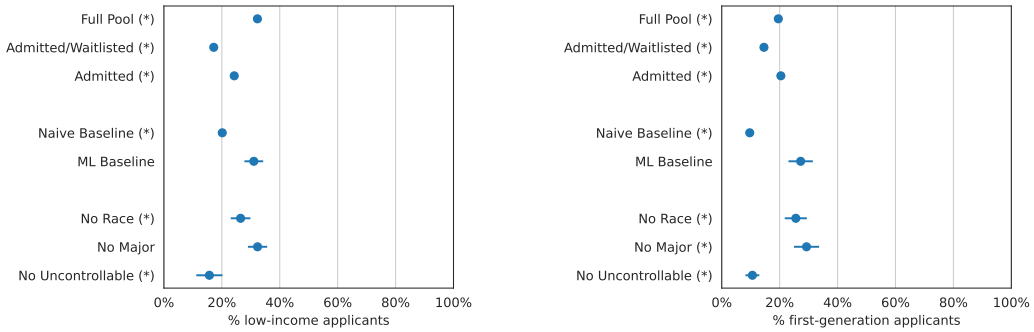
The reduction in diversity is not associated with a corresponding increase in academic merit of top-ranked applicants. Across all models, the academic merit of the top-ranked pool of applicants remains largely unchanged, as shown in Fig. 3. The average standardized test percentile (among applicants who submitted standardized test scores) of admitted applicants was 98.3, compared to the ML baseline average of 97.0. Excluding race from the ML baseline model results in a statistically significant ($p < 0.001$) increase in the average standardized test percentile of the top-ranked students, to 97.8. In absolute terms, however, this is a small change: it is approximately the difference between a 1480 and a 1490 on the SAT. Excluding data on applicant major preference does not meaningfully change standardized test percentiles of the top-ranked pool ($p = 0.74$); excluding all uncontrollable features has a similar impact to excluding race alone ($p < 0.001$).

In addition, other than the naive baseline (37%, $p < 0.001$), no model was practically different than the ML baseline at identifying students who were actually admitted or waitlisted by the case institution. About half of the students included in the top pool are actually admitted or waitlisted across the ML baseline (47%), no race (49%, $p = 0.003$), no major (46%, $p = 0.43$), and no uncontrollable features (46%, $p = 0.90$) models. **Overall, we predict that, if**



(a) (b)

Figure 1: Impact of policy changes on the racial and ethnic diversity of the top-rated group of applicants. Graph (a) shows the racial demographics of the applicant pool: the first three rows show demographics for the full, admitted or waitlisted, and admitted pools of applicants; the subsequent rows show the demographics of the *top* group of applicants under different ranking algorithms. Graph (b) shows the proportion of URM applicants in the top group under different ranking algorithms. In Graph (b), statistically significant differences in the proportion of URM applicants in the top-ranked group compared to the ML baseline are denoted with an asterisk. 95% confidence intervals for the ML models are shown based on results over 1,000 bootstraps.



(a) (b)

Figure 2: Impact of policy changes on the socioeconomic diversity of the top-rated group of applicants. Graphs (a) and (b) show the proportion of LI and FG applicants, respectively, in the top group under different ranking algorithms. Statistically significant differences in proportion of LI and FG applicants in the top-ranked group compared to the ML baseline are denoted with an asterisk. 95% confidence intervals for the ML models are shown based on results over 1,000 bootstraps.

admissions ranking algorithms are minimally modified to comply with the SFFA policy change, they will prioritize a less diverse, but not more academically meritorious, pool of applicants for review.[16]

4.2 Individual Impact

Inherent randomness in the modeling process leads to arbitrary outcomes, especially for top-ranked applicants. We calculated self-consistency for each applicant across 1,000 bootstraps of the ML baseline model, the cumulative distribution function of which is shown in Fig. 4(a). A self-consistency of 1 means that all (or none) of the bootstrapped models rank an applicant in the top pool, while a self-consistency of 0.5 means that exactly half of the models rank an applicant in the top pool. Most applicants have a relatively high self-consistency: across all applicants (blue curve), 31% of applicants have the highest possible self-consistency, and 69% of applicants have sc ≥ 0.95.[17] However, the ML baseline model more consistently identifies applicants who are *not* included in the top pool (green curve) than applicants who *are included* (orange curve).

Fig. 4(b) provides deeper insight into the consistency of individual applicant outcomes at sc ≥ 0.95. It shows that over two-thirds of applicants (69%) have consistent (or non-arbitrary)

[16] In Appendix E, we show that this prediction is robust to the specific definition of 'top' applicants.

[17] sc = 0.95 corresponds to agreement between 97.5% of models, a very high level of agreement.

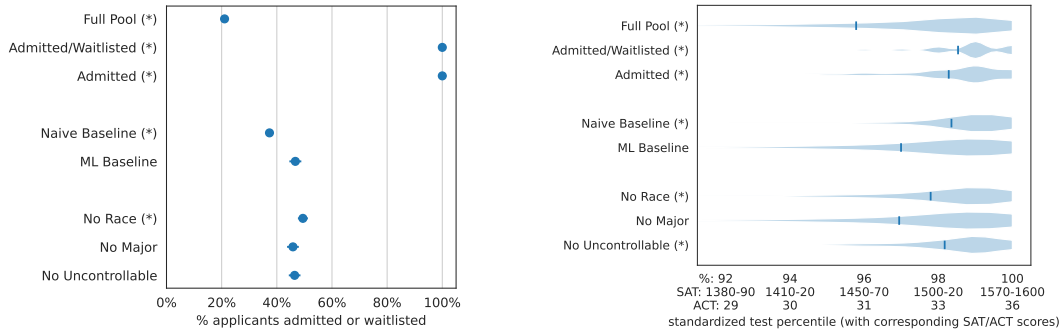


Figure 3: Impact of policy changes on the academic merit of the top-rated group of applicants. Graphs (a) and (b) show the proportion of actually admitted or waitlisted applicants and the distribution of standardized test score percentiles, respectively, in the top group under different ranking algorithms. Statistically significant differences in share of applicants actually admitted or waitlisted and standardized test percentile of the top group of applicants compared with the ML baseline are denoted with an asterisk. In Graph (a), 95% confidence intervals for the ML models are shown based on results over 1,000 bootstraps. In Graph (b), the darker blue line represents the *mean* standardized test percentile within the specified applicant pool.

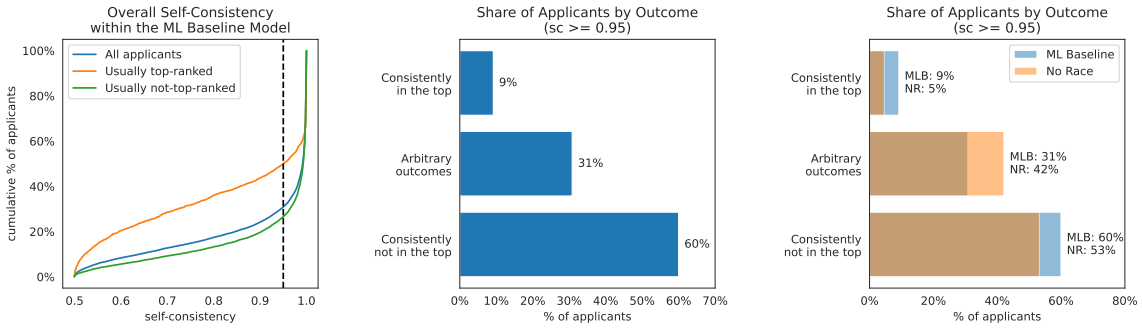


Figure 4: Graph (a) shows the cumulative distribution (CDF) of self-consistency within 1,000 bootstraps of the ML baseline model for the applicant pool (blue line), only applicants who are usually top-ranked (ranked in the top by >50% of bootstrapped models, orange line), and only applicants who are usually not top-ranked (ranked in the top by <=50% of bootstrapped models, green line). The dashed black line corresponds to sc = 0.95. Graph (b) shows the level of arbitrariness if we define an applicant's outcomes to be consistent if their sc $\geq$ 0.95 (and their outcomes to be arbitrary if their sc $<$ 0.95): only 9% of applicants are consistently ranked in the top, 60% of applicants are consistently not ranked in the top, and 31% of applicants have arbitrary outcomes. Graph (c) compares arbitrariness between the ML baseline and 'No race' (compliant with SFFA policy change) models.

outcomes. However, just 9% of these applicants are consistently ranked *in the top pool*, with the remaining 60% consistently ranked not in the top. Recall that the top pool consists of 20% of the applicant pool: this means that in any given bootstrapped model, more than half of the top pool consists of applicants who have been added to that pool somewhat arbitrarily. While Fig. 4(b) shows that this is the case for sc $\geq$ 0.95, Fig. 4(a) shows that this effect holds across all self-consistency thresholds. Across the entire distribution, a larger proportion of usually-top-ranked applicants (applicants who are ranked in the top pool by >50% of bootstrapped models) have lower self-consistencies than the usually-not-top-ranked applicants. Overall, we find that even within a single policy, randomness inherent to the modeling process has a major impact on who is included in the top pool.

Within-policy arbitrariness increases under the SFFA policy change. The ML baseline model is prohibited under the SFFA policy change because it explicitly considers applicants' race and ethnicity. A reasonable alternative is the 'no race' model, which is identical but does not consider race. Fig. 4(c) shows self-consistency within the ML baseline model (blue) and within the 'no race' policy (orange). Compared to the ML baseline model, the 'no race' model

has lower self-consistency – meaning that **inherent randomness, introduced through choices such as how to split training and test data, will play an even larger role in determining application review order following the SFFA policy change.** We provide a fuller analysis, including results that suggest that the SFFA policy change will increase arbitrariness for non-URM applicants in particular, in Appendix G.

5 Discussion

In this study, we investigated how changes in admission policies brought about by the end of affirmative action are likely to impact applicant ranking algorithms, which we argue can have a downstream impact on admissions decisions. We explored how hypothetical changes in admissions policies impact not only a college's overall class but also the outcomes of individual applicants. To do this, we predicted and contextualized the likely impact of race-unaware admissions on the ability of colleges to admit a first-year class that displays both academic merit and diversity, building on impactful prior work conducting demographic fairness assessments in college admissions [21, 28, 42, 59, 75]. We also provided a template for going beyond group fairness to understand the impact of policy changes on *individuals* by incorporating recent scholarship on model multiplicty and arbitrariness [16, 23].

We present three key findings. First, consistent with prior work [36], we find that **race-unaware policies decrease the proportion of URM applicants represented in the top-ranked pool by 62%**. Crucially, this change occurs **without a corresponding increase in academic merit of that top-ranked pool** (§4.1). Second, even in the absence of policy change, **inherent randomness in the modeling process will lead to somewhat arbitrary outcomes, especially for top-ranked applicants**: we find that across repeated bootstraps of the ML baseline model, just 9% of applicants are consistently ranked in the top 20% (§4.2). Third, **under a race-unaware applicant ranking algorithm, arbitrariness in individual outcomes increases relative to the baseline for most applicants** (§4.2).

In summary, our results imply that despite the impact of the SFFA policy change on college admissions processes, complaints long attributed to affirmative action will persist at highly selective institutions; for example, many students with high test scores may not be ranked highly or ultimately admitted. We propose that this is because those complaints stem not from the specifics of any policy, including affirmative action, but from the fundamental issues of (1) limited space at selective American colleges and (2) inherent randomness in the admissions process. Because these constraints are intrinsic to the admissions system, we argue that ending affirmative action will not resolve these issues.

5.1 Limitations

We acknowledge several limitations of our work. Chief among them is the narrow scope of our analysis. We focus on applicant ranking algorithms as one component of a larger admissions process and make an assumption that the order in which applicants are reviewed can impact admissions outcomes. While we believe that this is a reasonable assumption based on how admissions ranking algorithms have previously been implemented in practice [70], we are not able to precisely quantify this assumed impact. Further, by examining the impact of the SFFA policy change on ranking algorithms only, we tacitly accept much of the status quo of college admissions. For example, we choose to follow prior work [36] and train our applicant ranking algorithms on *past decisions*, effectively co-signing those as correctly identifying students with high academic merit (even though rejected students also may have had merit). In order to mitigate this, we conducted a robustness assessment of an alternative target variable specification (Appendix F), but we could have taken a more value-sensitive approach to applicant ranking algorithms instead (e.g. by exploring affirmative action based on socioeconomic status, as suggested by Chetty et al. [21]).

We also chose to accept the Common App's definition of 'under-represented minority candidate' and to assess diversity primarily according to that variable. Although URM status is an important element to consider in the admissions process, essential questions about the significance of race and ethnicity relative to other uncontrollable applicant features (e.g. legacy, FGLI status) remain. This focus is not merely specific to the institution in question but reflects broader societal and educational dynamics. Race often intersects with numerous other factors such as socioeconomic status, influencing higher educational opportunities and outcomes. Our study underscores the need to consider these intersections critically, recognizing race as a pivotal element in the complex matrix of college admissions.

More broadly, with this work, we focus on *computational solutions* to the *sociotechnical problem* of bias and inequity in college admissions and thus forgo an examination of more transformational changes that the SFFA policy change could inspire [6, 27]. However, we emphasize that our work is *not prescriptive*; instead, we have sought to measure and contextualize likely changes to the applicant review process resulting from the SFFA policy change. We hope that our findings—that the SFFA policy change is likely to decrease the share of URM candidates who are given priority reviewing without meaningfully increasing the predictive power of ranking algorithms to identify applicants with high academic merit—will inspire future work to substantively improve racial, socioeconomic, and other forms of diversity in higher education.

5.2 Opportunities for Future Work

By the time this work is published, data on the 2023-2024 college admissions cycle will be available, and researchers will be able to explore the extent to which our predictions on the impact of the SFFA policy change have come to pass, both in terms of how applicant ranking algorithms are modified and how the actually admitted class changes (and does not change). Early results are mixed – while many selective colleges have reported a decrease in URM enrollment following the SFFA policy change (and some have reported very large decreases) this is not universal [46]. We believe that it will be important to conduct an empirical validation of our results, including an analysis into where and why deviations from our predictions may occur (perhaps due to behavior changes from colleges and/or applicants in response to the SFFA policy change [29, 35, 40, 55] – for example, applicants could choose to disclose aspects of their identities in their essays, and colleges could choose to place more emphasis on elements of applicants' identities that they still have

access to under the SFFA policy change, such as socioeconomic status). In addition, our work suggests several avenues for future research. We show that, if previously built applicant ranking algorithms are minimally modified to be in compliance with the SFFA policy change, the share of top-ranked applicants who are URM is likely to fall by 62%. It will therefore be important to identify alternative ranking approaches that can mitigate this impact, or that can increase other forms of diversity, like the share of top-ranked applicants that are FGLI students. Prior work related to equity and access in algorithms, including Thomas et al. [67]'s framework for positive action, Arif Khan et al. [12]'s decision procedures for substantive equality of opportunity, and Borgs et al. [17]'s approach to algorithmic greenlining, will be instructive here; as will prior work on measuring and improving fairness without access to demographic data [14]. We also show that arbitrariness can have a major impact on how applicants are ranked by ML models. In addition to considering arbitrariness in individual outcomes as a component of fairness assessments going forward, future work could explore how to reduce this arbitrariness, perhaps through bagging as suggested by Cooper et al. [23] or through other variance reduction techniques.

6 Conclusion

In this work, we quantify how changes in the admissions process for selective American colleges – driven by a growing number of applications, test-optional policies, and the recent ban on race-conscious admissions – will impact the order in which applicants are prioritized for review, and, by extension, who is admitted. We find that the SFFA policy change is likely to reduce the share of top-ranked applicants who are URM without meaningfully increasing academic merit. Additionally, we find that inherent randomness in the modeling process will lead to somewhat arbitrary outcomes for individuals, especially for top-ranked applicants, and that arbitrariness is likely to increase as a result of the SFFA policy change.

Acknowledgments

This research was supported by NSF Grants (IIS-2008139, IIS-2312865), the Amazon Research Award, the Graduate Fellowships for STEM Diversity (GFSD), the Cornell Tech Urban Tech Hub Grant, and a seed grant from the Cornell Center for Social Sciences. The content represents the views of the authors and does not necessarily reflect the opinions of their respective employers or sponsors.

References

[1] 1978. University of California Regents v. Bakke. 438 U.S. 265.
[2] 2003. Gratz v. Bollinger. 539 U. S. 244.
[3] 2003. Grutter v. Bollinger. 539 U. S. 306.
[4] 2023. Students for Fair Admissions, Inc. v. President and Fellows of Harvard College. 600 U. S. ___.
[5] 2024. Report From Working Group on the Role of Standardized Test Scores in Undergraduate Admissions. Dartmouth College. https://home.dartmouth.edu/sites/home/files/2024-02/sat-undergrad-admissions.pdf
[6] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*. Association for Computing Machinery, New York, NY, USA, 252–260. https://doi.org/10.1145/3351095.3372871
[7] Columbia Undergraduate Admissions. [n. d.]. Testing Policy. https://undergrad.admissions.columbia.edu/apply/process/testing
[8] MIT Admissions. 2024. MIT Admissions. https://mitadmissions.org/
[9] Stanford Undergraduate Admissions. 2023. Admission Overview. https://admission.stanford.edu/apply/overview/index.html
[10] A.J. Alvero, Noah Arthurs, anthony lising antonio, Benjamin W. Domingue, Ben Gebre-Medhin, Sonia Giebel, and Mitchell L. Stevens. 2020. AI and Holistic Review: Informing Human Reading in College Admissions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. Association for Computing Machinery, New York, NY, USA, 200–206. https://doi.org/10.1145/3375627.3375871
[11] McKane Andrus and Sarah Villeneuve. 2022. Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, Seoul Republic of Korea, 1709–1721. https://doi.org/10.1145/3531146.3533226
[12] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2022. Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. ACM, Arlington VA USA, 1–10. https://doi.org/10.1145/3551624.3555303
[13] Richard Arum and Mitchell L. Stevens. 2023. For Most College Students, Affirmative Action Was Never Enough. The New York Times. https://www.nytimes.com/interactive/2023/07/03/opinion/for-most-college-students-affirmative-action-was-not-enough.html
[14] Carolyn Ashurst and Adrian Weller. 2023. Fairness Without Demographic Data: A Survey of Approaches. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3617694.3623234
[15] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing Fair Ranking Schemes. In *Proceedings of the 2019 International Conference on Management of Data* (Amsterdam, Netherlands) *(SIGMOD '19)*. Association for Computing Machinery, New York, NY, USA, 1259–1276. https://doi.org/10.1145/3299869.3300079
[16] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, Seoul Republic of Korea, 850–863. https://doi.org/10.1145/3531146.3533149
[17] Christian Borgs, Jennifer Chayes, Nika Haghtalab, Adam Tauman Kalai, and Ellen Vitercik. 2019. Algorithmic Greenlining: An Approach to Increase Diversity. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. Association for Computing Machinery, New York, NY, USA, 69–76. https://doi.org/10.1145/3306618.3314246
[18] William G. Bowen. 1977. Admissions and the Relevance of Race: Addressing the issues of principle, policy, and practice raised by the Bakke case. Princeton Alumni Weekly. https://www.princeton.edu/~paw/web_exclusives/more/article_archives_bowen.html
[19] Lilah Burke. 2020. The Death and Life of an Admissions Algorithm. Inside Higher Ed. https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd
[20] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. arXiv:1704.06840 [cs.DS]
[21] Raj Chetty, David Deming, and John Friedman. 2023. *Diversifying Society's Leaders? The Determinants and Causal Effects of Admission to Highly Selective Private Colleges*. Technical Report w31492. National Bureau of Economic Research, Cambridge, MA. w31492 pages. https://doi.org/10.3386/w31492
[22] Art Coleman and Jamie Lewis Keith. 2018. Understanding Holistic Review in Higher Education Admissions: Guiding Principles and Model Illustrations. College Board and EducationCounsel. https://highered.collegeboard.org/media/pdf/understanding-holistic-review-he-admissions.pdf
[23] A. Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. 2024. Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. arXiv:2301.11562
[24] Kuheli Dutt, Danielle L. Pfaff, Ariel F. Bernstein, Joseph S. Dillard, and Caryn J. Block. 2016. Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nature Geoscience* 9, 11 (Nov. 2016), 805–808. https://doi.org/10.1038/ngeo2819
[25] John Friedman, Bruce Sacerdote, and Michele Tine. 2024. Standardized Test Scores and Academic Performance at Ivy-Plus Colleges. Opportunity Insights. https://opportunityinsights.org/wp-content/uploads/2024/01/SAT_ACT_on_Grades.pdf
[26] Nikhil Garg, Hannah Li, and Faidra Monachou. 2021. Standardized Tests and Affirmative Action: The Role of Bias and Variance. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, New York, NY, USA, 261. https://doi.org/10.1145/3442188.3445889
[27] Ben Green. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology* 35, 4 (Dec. 2022), 90. https://doi.org/10.1007/s13347-022-00584-6

[28] Joshua Grossman, Sabina Tomkins, Lindsay Page, and Sharad Goel. 2023. *The Disparate Impacts of College Admissions Policies on Asian American Applicants*. Technical Report w31527. National Bureau of Economic Research, Cambridge, MA. w31527 pages. https://doi.org/10.3386/w31527

[29] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. 111–122.

[30] Anemona Hartocollis and Amy Harmon. 2023. Affirmative Action Ruling Shakes Universities Over More Than Race. The New York Times. https://www.nytimes.com/2023/07/26/us/affirmative-action-college-admissions-harvard.html

[31] Benjamin Q. Huynh, Elizabeth T. Chin, Allison Koenecke, Derek Ouyang, Daniel E. Ho, Mathew V. Kiang, and David H. Rehkopf. 2024. Mitigating allocative tradeoffs and harms in an environmental justice data tool. *Nature Machine Intelligence* 6, 2 (01 Feb 2024), 187–194. https://doi.org/10.1038/s42256-024-00793-y

[32] Common Data Set Initiative. 2023. Common Data Set 2023-2024. https://commondataset.org/wp-content/uploads/2024/01/CDS_2023-2024-with-auto-sum.pdf

[33] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[34] Brian Heseung Kim, Elyse Armstrong, Laurel Eckhouse, Mark Freeman, Rodney Hughes, Trent Kajikawa, and Michelle Sinofsky. 2024. Deadline updates, 2023–2024: First-year application trends through Feb 1. https://www.commonapp.org/files/Common-App-Deadline-Updates-2024.02.14.pdf

[35] Benjamin Laufer, Jon Kleinberg, Karen Levy, and Helen Nissenbaum. 2023. Strategic Evaluation. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3617694.3623237

[36] Hansol Lee, René F. Kizilcec, and Thorsten Joachims. 2023. Evaluating a Learned Admission-Prediction Model as a Replacement for Standardized Tests in College Admissions. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark) *(L@S '23)*. Association for Computing Machinery, New York, NY, USA, 195–203. https://doi.org/10.1145/3573051.3593382

[37] Jinsook Lee, Bradon Thymes, Joyce Zhou, Thorsten Joachims, and Rene F. Kizilcec. 2023. Augmenting Holistic Review in University Admission using Natural Language Processing for Essays and Recommendation Letters. In *Equity, Diversity, & Inclusion in Educational Technology Research & Development Workshop (AIED)*. arXiv, Tokyo, Japan. http://arxiv.org/abs/2306.17575

[38] David Leonhardt. 2023. Why Does Duke Have So Few Low-Income Students? The New York Times Magazine. https://www.nytimes.com/interactive/2023/09/07/magazine/duke-economic-diversity.html

[39] David Leonhardt. 2024. The Misguided War on the SAT. The New York Times. https://www.nytimes.com/2024/01/07/briefing/the-misguided-war-on-the-sat.html

[40] Lydia T Liu, Nikhil Garg, and Christian Borgs. 2022. Strategic ranking. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2489–2518.

[41] Zhi Liu and Nikhil Garg. 2021. Test-optional policies: Overcoming strategic behavior and informational gaps. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–13.

[42] Amelie Marian. 2023. Algorithmic Transparency and Accountability through Crowdsourcing: A Study of the NYC School Admission Lottery. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, New York, NY, USA, 434–443. https://doi.org/10.1145/3593013.3594009

[43] Kelly McConvey, Shion Guha, and Anastasia Kuzminykh. 2023. A Human-Centered Review of Algorithms in Decision-Making in Higher Education. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, Hamburg Germany, 1–15. https://doi.org/10.1145/3544548.3580658

[44] Katharine Meyer. 2023. The end of race-conscious admissions. The Brookings Institution. https://www.brookings.edu/articles/the-end-of-race-conscious-admissions/

[45] Kristen M. Glasener Michael N. Bastedo, Nicholas A. Bowman and Jandi L. Kelly. 2018. What are We Talking About When We Talk About Holistic Review? Selective College Admissions and its Effects on Low-SES Students. *The Journal of Higher Education* 89, 5 (2018), 782–805. https://doi.org/10.1080/00221546.2018.1442633

[46] James Murphy. 2024. Tracking the Impact of the SFFA Decision on College Admissions. Education Reform Now. https://edreformnow.org/2024/09/09/tracking-the-impact-of-the-sffa-decision-on-college-admissions/

[47] University of California Standardized Testing Task Force. 2020. Report of the UC Academic Council Standardized Testing Task Force. https://senate.universityofcalifornia.edu/_files/underreview/sttf-report.pdf

[48] U.S. Department of Education National Center for Education Statistics. [n. d.]. Integrated Postsecondary Education Data System (IPEDS), Admissions component final data (fall 2014 - 2021) and provisional data (fall 2022). https://nces.ed.gov/ipeds/TrendGenerator/app/answer/10/101

[49] U.S. Department of Education Office for Civil Rights. 2018. DATA HIGHLIGHTS ON SCIENCE, TECHNOLOGY, ENGINEERING, AND MATHEMATICS COURSE TAKING IN OUR NATION'S PUBLIC SCHOOLS. https://www2.ed.gov/about/offices/list/ocr/docs/stem-course-taking.pdf 2015–16 CIVIL RIGHTS DATA COLLECTION STEM COURSE TAKING.

[50] General Assembly of the State of Indiana. 1843. *The revised statutes of the state of Indiana, passed at the twenty-seventh session of the General assembly*. J. Dowling and R. Cole, state printers, Indianapolis. 1154 pages. https://babel.hathitrust.org/cgi/pt?id=nyp.33433009073879&seq=352 By an act of the General assembly Samuel Bigger was 'authorized to prepare a compilation and revision of the general statute laws'. George H. Dunn was associated with the reviser.

[51] General Assembly of the State of North Carolina. 1830. Acts passed by the General Assembly of the State of North Carolina (1830-1831). https://digital.ncdcr.gov/Documents/Detail/acts-passed-by-the-general-assembly-of-the-state-of-north-carolina-1830-1831/1955764?item=2080405 Chapter VI: A Bill to Prevent All Persons from Teaching Slaves to Read or Write, the Use of Figures Excepted (page 15).

[52] Presidential Committee on Harvard & the Legacy of Slavery, Tomiko Brown-Nagin, Sven Beckert, Annette Gordon-Reed, Stephen Gray, Evelynn M. Hammonds, Nancy F. Koehn, Meira Levinson, Tiya Miles, Martha Minow, Maya Sen, Daniel Albert Smith, David R. Williams, and William Julius Wilson. 2022. Harvard & the Legacy of Slavery. https://legacyofslavery.harvard.edu/

[53] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[54] Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair ranking: a critical review, challenges, and future directions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 1929–1942.

[55] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 7599–7609. https://proceedings.mlr.press/v119/perdomo20a.html

[56] Abdul Hamid M Ragab, Abdul Fatah S Mashat, and Ahmed M Khedra. 2012. HRSPCA: Hybrid recommender system for predicting college admission. In *2012 12th International conference on intelligent systems design and applications (ISDA)*. IEEE, 107–113.

[57] S.F. Reardon. 2011. The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. https://cepa.stanford.edu/sites/default/files/reardon%20whither%20opportunity%20-%20chapter%205.pdf In R. Murnane & G. Duncan (Eds.), Whither Opportunity? Rising Inequality and the Uncertain Life Chances of Low-Income Children. New York: Russell Sage Foundation Press.

[58] Sarah Reber, Gabriela Goodman, and Rina Nagashima. 2023. Admissions at most colleges will be unaffected by Supreme Court ruling on affirmative action. The Brookings Institution. https://www.brookings.edu/articles/admissions-at-most-colleges-will-be-unaffected-by-supreme-court-ruling-on-affirmative-action/

[59] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling Assumptions Clash with the Real World: Transparency, Equity, and Community Challenges for Student Assignment Algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445748

[60] Lucy Shao, Richard A. Levine, Stefan Hyman, Jeanne Stronach, and Juanjuan Fan. 2022. A Combinatorial Optimization Framework for Scoring Students in University Admissions. *Evaluation Review* 46, 3 (2022), 296–335. https://doi.org/10.1177/0193841X221082887

[61] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088

[62] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf

[63] Sashank Sridhar, Siddartha Mootha, and Santosh Kolagati. 2020. A university admission prediction system using stacked ensemble learning. In *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*. IEEE, 162–167.

[64] Shawn Staudaher, Jeonghyun Lee, and Farahnaz Soleimani. 2020. Predicting Applicant Admission Status for Georgia Tech's Online Master's in Analytics Program. In *Proceedings of the Seventh ACM Conference on Learning @ Scale* (Virtual Event, USA) *(L@S '20)*. Association for Computing Machinery, New York, NY, USA, 309–312. https://doi.org/10.1145/3386527.3406735

[65] Mitchell L. Stevens. 2007. *Creating a Class: College Admissions and the Education of Elites.* Harvard University Press. http://www.jstor.org/stable/j.ctv1m46g11
[66] Lisa M. Stulberg and Anthony S. Chen. 2014. The Origins of Race-conscious Affirmative Action in Undergraduate Admissions: A Comparative Analysis of Institutional Change in Higher Education. *Sociology of Education* 87, 1 (2014), 36–52. https://doi.org/10.1177/0038040713514063
[67] Oliver Thomas, Miri Zilka, Adrian Weller, and Novi Quadrianto. 2021. An Algorithmic Framework for Positive Action. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. ACM, – NY USA, 1–13. https://doi.org/10.1145/3465416.3483303
[68] Carnegie Mellon University. 2024. Admissions Consideration. https://www.cmu.edu/admission/admission/admission-consideration
[69] Dineshkumar B Vaghela and Priyanka Sharma. 2015. Students' Admission Prediction using GRBST with Distributed Data Mining. *Communications on Applied Electronics* 2, 1 (2015), 15–19.
[70] Austin Waters and Risto Miikkulainen. 2014. GRADE: Machine-Learning Support for Graduate Admissions. *AI Magazine* 35, 1 (March 2014), 64–75. https://doi.org/10.1609/aimag.v35i1.2504
[71] Jamelle Watson-Daniels, Solon Barocas, Jake M. Hofman, and Alexandra Chouldechova. 2023. Multi-Target Multiplicity: Flexibility and Fairness in Target Specification under Resource Constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, New York, NY, USA, 297–311. https://doi.org/10.1145/3593013.3593998
[72] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1569–1578. https://doi.org/10.1145/3132847.3132938
[73] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part I: Score-Based Ranking. *ACM Comput. Surv.* 55, 6, Article 118 (dec 2022), 36 pages. https://doi.org/10.1145/3533379
[74] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems. *ACM Comput. Surv.* 55, 6, Article 117 (dec 2022), 41 pages. https://doi.org/10.1145/3533380
[75] Anna Zimdars. 2010. Fairness and undergraduate admission: a qualitative exploration of admissions choices at the University of Oxford. *Oxford Review of Education* 36, 3 (June 2010), 307–323. https://doi.org/10.1080/03054981003732286

A Ethical Considerations

Adverse Impact. Our aim with this work is to measure and contextualize challenges to admitting a diverse first-year class with high academic merit resulting from the SFFA policy change. However, we acknowledge that, by highlighting the modeling changes that most reduce the share of top-ranked URM, FG, and LI candidates, our work has the potential to be misused by those who wish to *reduce* diversity in college admissions. However, because the changes we highlight are relatively simple (i.e. removal of sensitive variables), we argue that they are not likely to reveal previously unknown methods of discrimination, and therefore we believe the value in understanding the impact of the SFFA policy change outweighs the risk.

Privacy and Human Subjects Research. As part of this study, we had access to personal and sensitive information from college applicants. We took care to protect these data throughout the process of this study. All data was stored securely on dedicated servers that could only be accessed by approved individuals. All personally identifying information, including applicant names and contact information, were removed before we accessed and analyzed the data. This study was determined to be exempt by our institution's IRB.

B Admissions at Selective American Colleges

C Preprocessing: Additional Details

To understand how the SFFA policy change is likely to impact *already existing* applicant ranking algorithms, we followed similar data acquisition and preprocessing steps to those already implemented by the case institution. We included all data that was common across all four years of applications, except for personally identifiable information (name, date of birth, contact information, etc.) and data that is not entered directly into the Common App form but is instead provided as a file upload (transcript, letters of recommendation, etc.). Ultimately, this left us with 302 raw features; as part of preprocessing (following a similar approach to that outlined in [36]), we also applied a one-hot encoding to categorical features, recoded categories that occur in fewer than 1% of observations as 'RARE', imputed missing values and added indicator variables indicating that a numeric feature was missing (categorical variables were directly imputed as 'MISSING'), and constructed TF-IDF unigrams and bigrams for text features. We split our data into train and test sets based on year: we used the 2019-2020, 2020-2021, and 2021-2022 admissions cycles as training data and the 2022-2023 admissions cycle as test data. This mimics the real-world scenario of training a model on all available historical data, and also allows us to use results for a complete applicant pool as test data. Summary statistics describing our training and test data sets are presented in Table 1.

D Defining Baselines: Additional Details

To measure the impact of the SFFA policy change, we first need to establish a baseline. Here we rely on two baselines. First, we train an *ML baseline* model that uses every variable available from the processed Common App data (as described above) to predict an applicant's likelihood of being accepted. Following Lee et al. [36], we

Table 2: Details of the admissions process at Ivy-Plus institutions (which we define following Chetty et al. [21]). All data comes from the most recently released version of the Common Data Set for each institution.

Institution	# Applicants	# Admitted	# Waitlist	Application Due Date	Notification Date
Brown University	50,649	2,562	-	January 5	Late March
Columbia University	60,374	2,255	-	January 1	April 1
Cornell University	71,164	5,168	7,729	January 2	Early April
Dartmouth College	28,336	1,808	2,098	January 3	Early April
Duke University	49,523	2,911	-	January 3	April 1
Harvard University	61,221	1,984	-	January 1	April 1
MIT	33,767	1,337	763	January 1	March 20
Princeton University	38,019	2,167	1,710	January 1	April 1
Stanford University	56,378	2,075	553	January 5	April 1
University of Chicago	37,974	2,460	-	-	-
University of Pennsylvania	54,588	3,549	3,351	January 5	April 1
Yale University	50,060	2,289	1,000	January 2	April 1

Sources: https://oir.brown.edu/sites/default/files/2020-04/CDS_2022_2023.pdf,
https://opir.columbia.edu/sites/default/files/content/Common%20Data%20Set/CDS%20College%20Engineering%202022-2023.pdf,
https://irp.dpb.cornell.edu/wp-content/uploads/2023/09/CDS_2022-2023_Cornell-University-v7.pdf, https://www.dartmouth.edu/oir/pdfs/cds_2022-2023.pdf,
https://provost-files.cloud.duke.edu/sites/default/files/CDS%202021-22%20FINAL_2.pdf,
https://bpb-us-e1.wpmucdn.com/sites.harvard.edu/dist/6/210/files/2023/06/harvard_cds_2022-2023.pdf, https://ir.mit.edu/cds-2023,
https://registrar.princeton.edu/sites/g/files/toruqf136/files/documents/CDS_2022-2023.pdf,
https://ucomm.stanford.edu/wp-content/uploads/sites/15/2023/03/CDS_2022-2023_v3.pdf,
https://bpb-us-w2.wpmucdn.com/voices.uchicago.edu/dist/8/2077/files/2022/10/UChicago_CDS_2021-22.pdf, https://upenn.app.box.com/s/75jr7yip7279rcsfic0946o1pkr8okrt,
https://oir.yale.edu/sites/default/files/cds_yale_2022-2023_vf_10062023.pdf

believe that this model is a reasonable representation of how college admissions offices might incorporate ML into their processes.

We also define a *naive baseline* model that ranks applicants based first on their highest level of prior math instruction and then on their standardized test scores. Lee et al. [36] suggest that relying on standardized test scores represents an applicant ranking method that was commonly used pre-COVID. Because our test set contains data from the 2022-2023 admissions cycle (i.e. after test-optional policies were put in place at our case institution), we are missing standardized test scores for 32% of the applicant pool. Therefore, we supplement our baseline ranking with data on applicants' highest level of math taken, which is both salient to our (engineering-focused) case institution and available for the entire pool of applicants. We first rank applicants according to their highest math course taken.[18] Next, we convert applicants' reported SAT[19] and ACT[20] scores to percentiles to make them directly comparable with one another. Then, within the band of 'highest math taken,' we rank applicants according to the higher of their two percentiles (applicants who did not report either score are ranked last within their band[21]). Finally, we again select the top pool of applicants where we assume an admissions office would focus the majority of their attention. Mirroring the ML approach described above, we define this as the set of applicants with the top 20% of baseline scores.[22]

E Robustness to Different 'Top Pool' Cutoffs

To ensure that our results are not brittle with respect to our specific choice of how to define the 'top' pool of applicants, we conducted a robustness assessment, examining to what extent our findings about diversity and academic merit of the top-ranked pool change as the top pool itself changes. To do this, we vary the 'cutoff' for the top pool: the minimum decile considered part of the top. As Fig. 5 shows, the relative ordering of attributes across models remains constant at almost all cutoff choices. For example, the simulated policy changes that result in a decreased URM share within the top pool of applicants relative to the ML baseline model (the 'no race' and 'no uncontrollable features' models) do so whether the cutoff is set at Decile 10, Decile 9, Decile 8, Decile 7, and so on. The only exception to this is that the share of FG applicants included in the top pool by the 'no race' model *increases* relative to the ML baseline if the cutoff is set at Decile 10, but *decreases* relative to the ML baseline if the cutoff is set at any other decile. The results we discuss in §4 use Decile 9 as the cutoff, which is consistent with most other cutoffs. Additionally, the magnitude of differences between models decreases as the cutoff decile decreases. This makes intuitive sense, as it indicates that a higher proportion of the overall applicant pool is included in the 'top' pool. When the cutoff is Decile 1, all applicants are included in the top pool, and the demographic and

[18]We note that, by ranking students according to their highest math class taken, the baseline model likely penalizes URM students for factors beyond their control. The U.S. Department of Education's Office for Civil Rights reports that just 38% of public high schools with high (≥75%) Black and Hispanic enrollment offer calculus, compared with 50% of public high schools nationally [49]. We do not suggest this model be used to rank applicants in practice and only put it forward as a baseline.

[19]https://research.collegeboard.org/reports/sat-suite/understanding-scores/sat

[20]https://www.act.org/content/act/en/products-and-services/the-act/scores/national-ranks.html

[21]How to rank students who do not report a test is a non-trivial choice due to informational differences, fairness concerns to both those who report and do not report scores, and concerns about strategic reporting behavior [41]. However, because we consider test scores only within a math course band, the effect of this choice is relatively small.

[22]However, we note that because this metric is coarser than our predictive model, there are a large number of ties among top-scoring applicants. Taking ties into account, 21% of applicants share the top 20% of scores.

academic features of the top applicants converge to the averages of the test set.

F Robustness to Target Variable Selection

The choice of target variable can have a major impact on the outputs of applicant ranking algorithms. To ensure that our results are not brittle with respect to our specific choice of target variable, we examined our group impact findings for both diversity and academic merit to determine whether our results still hold when ML ranking algorithms are trained to predict applicants' likelihood of being *admitted or waitlisted* instead of simply being *admitted*. This is an important choice, with potentially significant implications. The decision to admit an applicant in a holistic admissions process is based on a variety of qualitative and quantitative factors, including that applicant's background. To the extent that prior admissions decisions are based on affirmative action, it could arguably be a violation of the SFFA policy change to train admissions algorithms on prior admissions decisions, because such a practice would effectively encode and carry forward affirmative action. However, affirmative action may have less of an impact on who is placed on the waitlist. We see empirical evidence for this in Fig. 6(a), which shows that the racial demographics of applicants who are admitted or waitlisted are more similar to the demographics of the full applicant pool than the racial demographics of applicants who are admitted. Therefore, training ranking algorithms on waitlist decisions may represent a reasonable course of action for admissions offices, and we explore its impacts.

As expected the ML baseline that is trained to identify admitted or waitlisted applicants includes a lower share of URM, LI, and FG students in the top pool. It also includes in the top pool applicants with slightly higher standardized test scores. However, the trends discussed in §4 still hold. Excluding race variables decreases the share of URM applicants in the top pool relative to the ML baseline Fig. 6(b); it also reduces the share of LI applicants Fig. 6(c) but does not meaningfully change the share of FG applicants Fig. 6(d). At the same time, all ML models identify similar shares of applicants who were actually admitted or waitlisted in their top pools Fig. 6(e), and they also identify applicants with similarly high standardized test scores Fig. 6(f).

G Comparing Arbitrariness Across Policies

Comparing sources of arbitrariness. To understand how much a policy change impacts an individual applicant's outcomes as compared to inherent randomness for a given policy, we compare self-consistency in applicants' outcomes *across* different policies to their self-consistency *within* bootstrapped instances of a single policy. To create a set of within-policy models M_{within}, we train one model on 1,000 bootstrapped samples of the training data. To create a set of across-policy models M_{across}, we sample 500 instances from the M_{within} set of one model and 500 instances from the M_{within} set of another model, creating a set of 1,000 modeling outcomes that represents two different policies. To compare the overall level of arbitrariness (the complement of self-consistency) *across* vs. *within* policies, we calculate the *arbitrariness ratio* AR:

$$\text{AR} = \frac{1 - \bar{\text{SC}}_{across}}{1 - \bar{\text{SC}}_{within}} \tag{2}$$

Intuitively, this is the ratio of the average level of arbitrariness across policies to the average level of arbitrariness within the ML baseline policy. The idea is that arbitrariness resulting from bootstrapping will be captured by both the M_{within} and M_{across} models, and any additional arbitrariness in the M_{across} models will be attributable to policy change. We also test whether the overall level of arbitrariness across all applicants is statistically significantly different across vs. within policies with a Wilcoxon signed-rank test, again with the standard $p = 0.05$ threshold.

Within-policy arbitrariness increases under the SFFA policy change. The ML baseline model is prohibited under the SFFA policy change because it explicitly considers applicants' race and ethnicity as a feature in prioritizing them for review. A reasonable alternative to the ML baseline model is the 'no race' model, which is identical but for the fact that it does not consider race as a feature. Fig. 7(a) shows self-consistency within the ML baseline model (blue curve) and within the 'no race' policy (orange curve). Compared to the ML baseline model, the 'no race' model has lower self-consistency – meaning that applicants may be even more susceptible to experiencing arbitrary outcomes as a result of inherent randomness after the SFFA policy change takes effect. Inherent randomness in the 'no race' modeling process creates arbitrariness that is 39% larger than the arbitrariness created by inherent randomness in the ML baseline modeling process (AR = 1.39; Wilcoxon signed-rank test: $p < 0.001$).

To provide a concrete example, we again compare consistency in outcomes at SC ≥ 0.95 in Fig. 7(c). Under the 'no race' model, 58% of applicants have consistent outcomes (compared to 69% under the ML baseline model), and 5% of applicants are consistently ranked in the top pool (compared to 9% under the ML baseline model). As Fig. 7(a) shows, this reduced consistency within the 'no race' model holds across all self-consistency thresholds. Again recalling that the top-ranked pool consists of 20% of the full applicant pool, this means that under the 'no race' model, three-quarters of the top pool will consist of applicants who have been added to that pool somewhat arbitrarily. Overall, these results imply that **inherent randomness, introduced through choices such as how to split training and test data, will play an even larger role in determining application review order following the SFFA policy change.**

Arbitrariness across policies is larger than arbitrariness within a single policy. Fig. 7(a) further shows self-consistency across both the ML baseline model and the 'no race' policy (green curve). We observe that while the 'no race' policy exhibits more arbitrariness than the ML baseline, the arbitrariness across these two policies is even larger. Intuitively, this means that changing a policy (in this case, to comply with the SFFA ruling), changes the outcomes that an applicant has, *even though their overall merit as an applicant does not change.* To quantify this more precisely, the overall arbitrariness ratio AR of the across-policy outcomes to outcomes within the ML baseline model is 1.66. This means that the policy change creates a level of arbitrariness that is 66% higher than the inherent randomness present in the ML baseline modeling process alone (Wilcoxon signed-rank test: $p < 0.001$). The AR of the across-policy outcomes to outcomes within the 'no race' model is 1.19 – the policy change increases arbitrariness ($p < 0.001$), beyond the 'no race' model. The observed pattern is even more pronounced for usually-top-ranked applicants, as shown in Fig. 7(b). For these

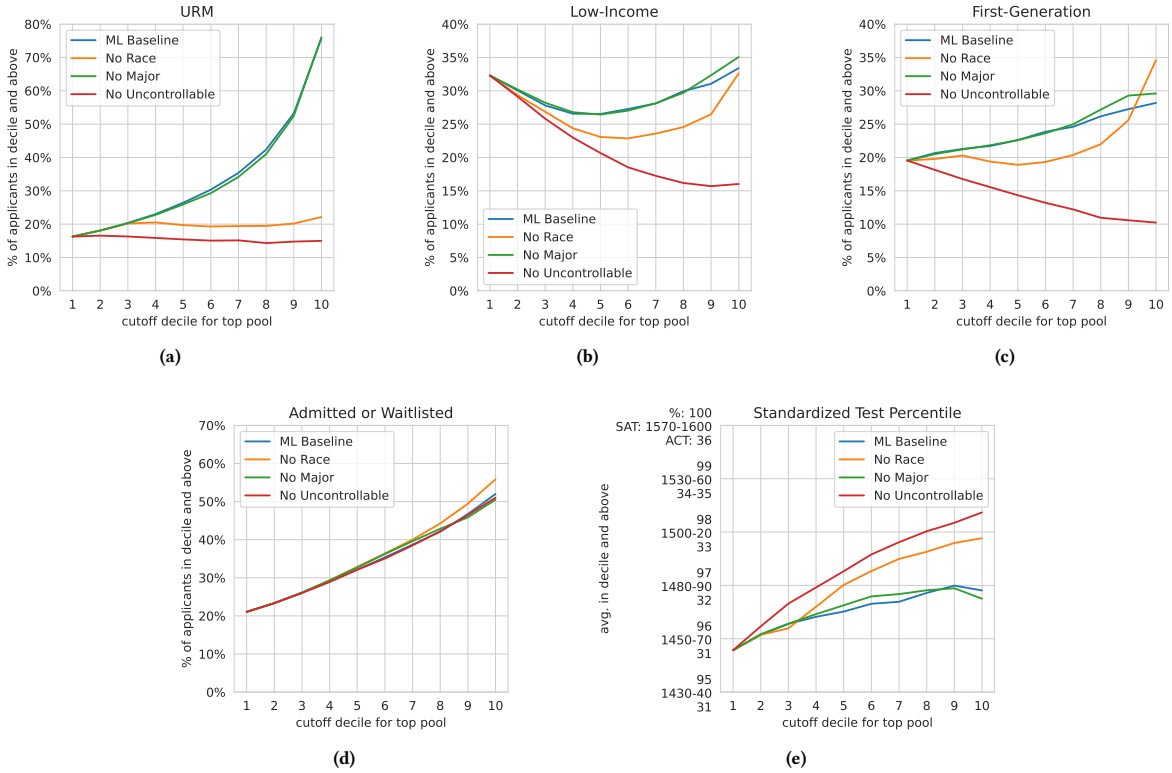


Figure 5: Graphs (a), (b), (c), and (d) show the share of applicants in the top pool who belong to each of the specified groups as the 'cutoff' (the minimum decile considered part of the top group) changes. Graph (e) shows the average standardized test percentile submitted by applicants who belong to the top group as the cutoff changes. Note that a cutoff value of 9 corresponds to the top group as defined in §4.

applicants, the arbitrariness ratios of across-policy outcomes to outcomes within the ML baseline and 'no race' models are 1.86 and 1.26, respectively (both $p < 0.001$). This means that the increase in arbitrariness created by policy change is *even higher* among applicants who are usually top-ranked.

Within-policy arbitrariness increases for specific groups of applicants under the SFFA policy change. Finally, we examine if the within-policy pattern of arbitrariness holds across racial and ethnic groups: as shown in Fig. 8, arbitrariness in outcomes statistically significantly increases for Hispanic, Native Hawaiian/Other Pacific Islander, multiracial, Asian, and White applicants, as well as applicants who did not report their race. Arbitrariness slightly decreases for Black applicants;[23] due to the under-representation of American Indian/Alaska Native applicants in the pool, we are not able to observe statistically significant changes in arbitrariness for that group.

[23]As implied by Fig. 1(a), this is due to the fact that Black applicants are more consistently *not* ranked in the top under the 'no race' model.

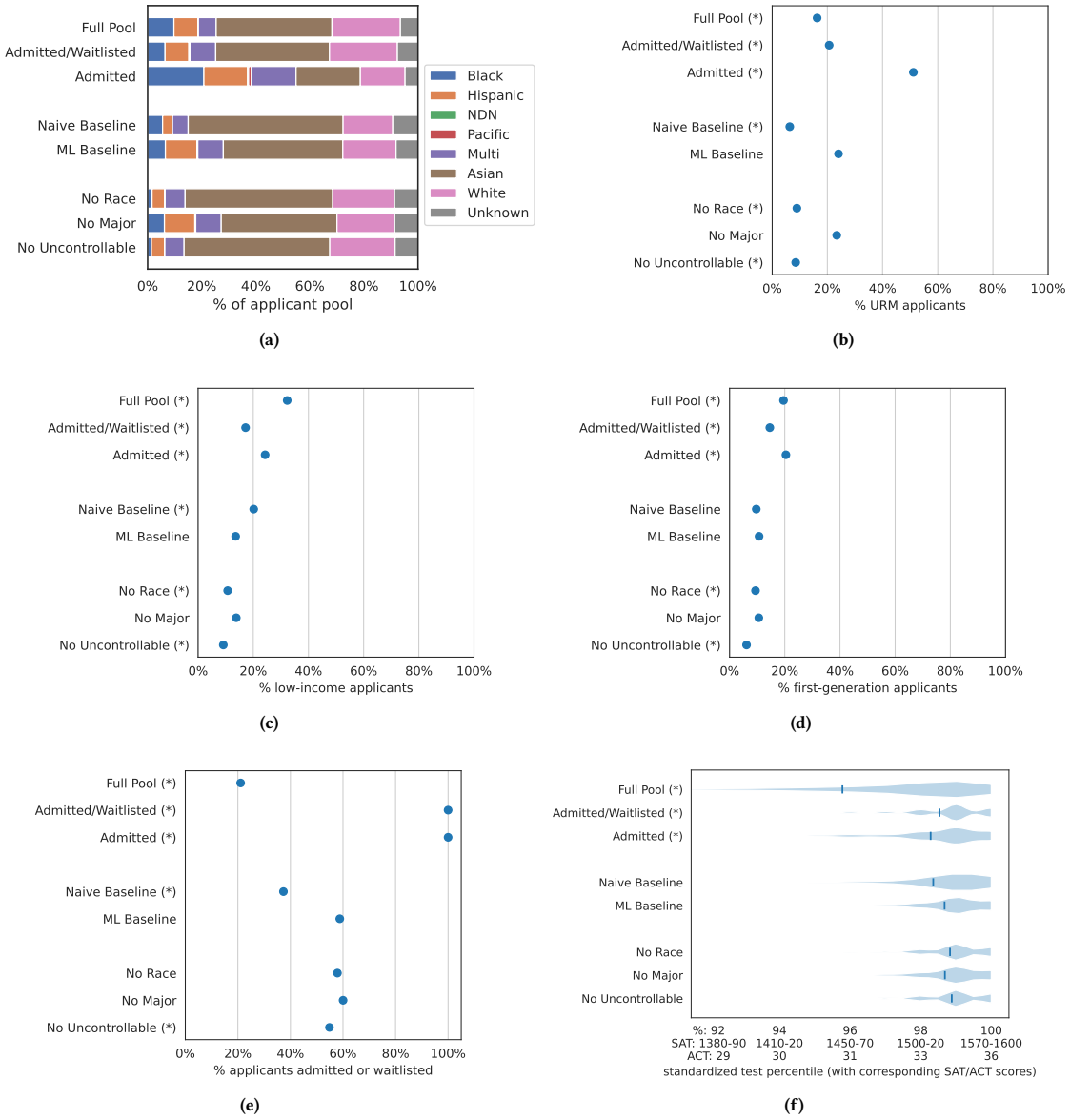


Figure 6: Group outcomes when the ML ranking algorithms are trained to predict applicants' likelihood of being *accepted or waitlisted* instead of only being *accepted*.

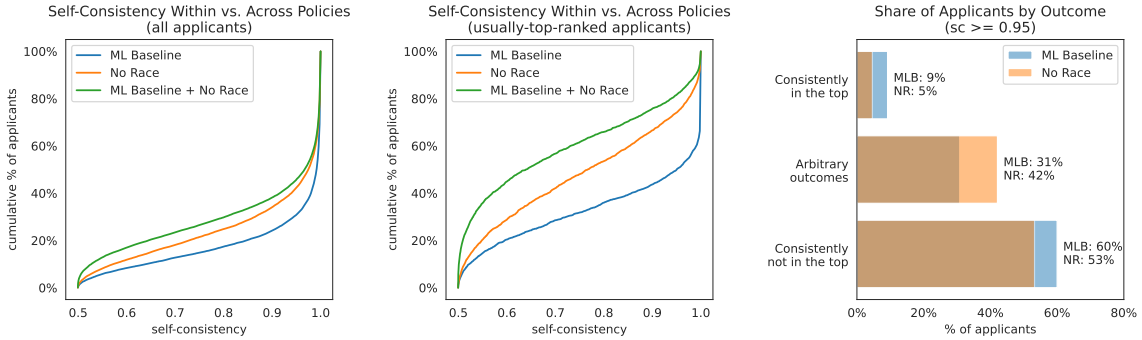


Figure 7: Graph (a) shows the CDF of self-consistency for all applicants within 1,000 bootstraps of the ML baseline model for the applicant pool (blue line), within 1,000 bootstraps of the 'no race' model (orange line), and across 500 bootstraps of each models (green line). Graph (b) shows CDFs only for those applicants who are usually top-ranked (ranked in the top by >50% of bootstrapped models). Graph (c) shows the level of arbitrariness in the ML baseline model compared to the 'no race' model, if we define an applicant's outcomes to be consistent if and only if their sc $\geq$ 0.95: 9% of applicants are consistently ranked in the top under the ML baseline model, and 5% are consistently ranked in the top under the 'no race' policy.

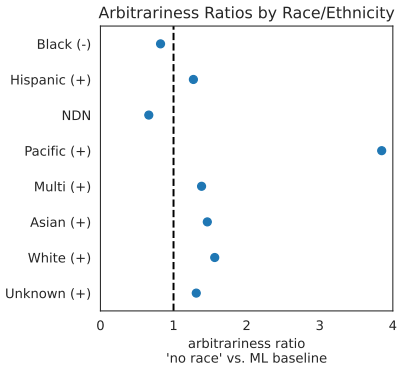


Figure 8: The arbitrariness ratios between the 'no race' and the ML baseline policies for applicants of different races and ethnicities. A '+' denotes that arbitrariness for the group statistically significantly *increases*, while a '-' denotes that arbitrariness for the group statistically significantly *decreases*, per a Wilcoxon signed-rank test with the Benjamini-Hochberg procedure applied ($p < 0.001$ in all cases, for except Native Hawaiian/Other Pacific Islander applicants, for whom $p = 0.04$).