Title
Qualitative information in undergraduate admissions: A pilot study of letters of recommendation

Author
Rothstein, Jesse

Publication Date
2022-07-28

Peer reviewed

Economics of Education Review 89 (2022) 102285

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Qualitative information in undergraduate admissions: A pilot study of letters of recommendation

Jesse Rothstein [1]

University of California, Berkeley, United States

ARTICLE INFO

JEL:
I21
I23
I25

Keywords:
College admissions
Natural language processing
Holistic review

ABSTRACT

A subset of undergraduate applicants to the University of California, Berkeley were invited to submit letters of recommendation as part of their applications. I use scraped text of the submitted letters, natural language processing tools, and a within-subject experimental design wherein applications were read in parallel with and without their letters to understand the role that this qualitative information plays in admissions. I show that letters written on behalf of underrepresented applicants were modestly distinctive. I also construct an index of letter strength, measuring the predicted impact of the letter on the student's application score. I show that underrepresented applicants tend to get weaker letters, but that readers pay less attention to letter strength for underrepresented students. Overall, the inclusion of letters modestly improved application outcomes for the average underrepresented student.

1. Introduction

The college admissions decision is an important determinant of students' life chances, and a highly visible source of inequality of opportunity. Students from disadvantaged backgrounds are much less likely to apply to (Hoxby & Avery, 2013; Hoxby & Turner, 2013) or be accepted at (Krueger et al., 2006) selective colleges than are those from more advantaged backgrounds. Growing evidence indicates that this has important implications for graduation and other life outcomes (e.g., Black, Denning, & Rothstein, 2021; Bleemer, 2021; Cohodes & Goodman, 2014).

The specific criteria used in making admissions decisions may have implications for the equity of the process. In particular, critics of the SAT and ACT college admissions exams have often argued that they create unfair advantages for students from already advantaged backgrounds (e.g., Atkinson, 2001; Kurlaender et al., 2020; Rothstein, 2004). This has motivated many colleges to change the weight placed on different quantitative measures, for example by emphasizing class rank, and others to adopt "holistic" review processes that aim to reduce reliance on quantitative measures more generally. Evidence on so-called "percent plans" suggests that merely changing the weights can lead to fairer admissions rules (Black, Denning, & Rothstein, 2021; Bleemer, 2021), though it does not fully offset inequality in pre-college opportunity (Cortes & Klasik, 2020).

But for all of the concern about the SAT, the qualitative components of college applications may be at least as much of a problem (Alvero et al., 2020, 2021). Students from lower-income and otherwise disadvantaged backgrounds are unlikely to get the college counseling, essay writing help, and other application assistance that is routinely available to more privileged students. This may lead them to submit weaker applications, with less polished or less carefully chosen essays. They may also have limited access to recommenders who can write them strong letters that satisfy the particular demands of undergraduate admissions offices. More generally, the inclusion of qualitative measures typically requires using subjective application reviews rather than mechanical formulas. It is not clear that more subjective decisions will be any fairer than are purely quantitative measures. Human reviewers are subject to their own biases, conscious or otherwise, and holistic review may simply expand the scope for them to preference students from more advantaged backgrounds.

[1] rothstein@berkeley.edu 2607 Hearst Avenue, Berkeley CA 94720-7320. I thank the UC Berkeley admissions office and administration for help in executing this project, particularly Carol Christ, Catherine Koshland, Greg Dubrow, Amy Jarich, and Olufeme Ogundole. The project was carried out with the support of the California Policy Lab, particularly Elsa Augustine, Charles Davis, Patrick Kennedy, Karla Palos, and Audrey Tiew. I thank Avi Feller, Mitchell Stevens, and Evan White for helpful conversations. I am grateful to the William T. Grant Foundation for financial support. The work was carried out while I was a faculty member at the University of California, Berkeley, the institution under study, and the study design was developed in consultation with the UC Berkeley admissions office and academic leadership. No one other than me had input into the results.

https://doi.org/10.1016/j.econedurev.2022.102285
Received 2 May 2022; Accepted 2 July 2022
Available online 28 July 2022

I shed light on this question by examining the role of letters of recommendation (LORs) in a selective campus's holistic admissions review process. I use natural language processing methods to measure the content of letters, and investigate whether letters submitted for students from different groups are distinctive (that is, whether the student's identity characteristics are reliably predicted by the letter content) and whether they improve or hurt those students' relative chances of admission. In contrast to past work using artificial intelligence to assess qualitative components of college applications (e.g., Alvero et al., 2020, Alvero et al., 2021), I do not make assumptions about how application reviewers interpret information that may be correlated with student background, but use actual application scores to infer what the readers value.

My study draws on a pilot program implemented by University of California, Berkeley (UCB) in 2016–17. Importantly, UCB is officially committed to an equitable admissions process that does not provide unfair advantages for students from advantaged backgrounds. It has for many years used "holistic review," whereby applicants are evaluated as a whole package without fixed weights for any particular admissions measures (Hout, 2005).[2] Applications are read individually, and readers are instructed not to use mechanical rules and to look for evidence that students from less advantaged backgrounds can succeed. However, prior to the pilot I study, UCB had never solicited letters of recommendation as part of its admissions process. The explicit motivation for introducing LORs was to provide more information that could be used to identify students who had overcome disadvantages, and the letters may have been used differently than in other settings where that is less of a priority.

Even within the UCB context the move to use LORs was controversial. While some, including the university's leadership, believed that LORs could be used to identify more students from underrepresented groups who were prepared to succeed at Berkeley, others argued that they would have the opposite effect. They predicted that low-income and otherwise disadvantaged applicants would not have access to mentors who could write them strong letters, and that the effect of including this qualitative information would be to create barriers to the admission of underrepresented applicants.[3]

The LOR pilot was designed to ask for letters only when they might be important to the decision because the applicant was otherwise marginal for admission. This created variation across students in the presence of letters. Ben-Michael, Feller, & Rothstein, 2021 study the effect of including LORs on admissions decisions at Berkeley, using matching strategies to identify comparable groups of students who did and did not submit letters. They find that the availability of letters improved application scores on average for both advantaged and disadvantaged groups, with the largest effects for students whose quantitative qualifications would have predicted modest probabilities of admission. The net effect was to slightly reduce the relative scores of students from underrepresented groups.

Here, I take advantage of a within-subjects experimental design that was built into the UCB pilot study to dig deeper into the roles that letters play. A subset of 10,000 applicants who had submitted letters was selected at random for inclusion in the experiment. These students' applications were evaluated twice – once as intended under the LOR pilot, with the letters considered by application readers, and once by a separate set of readers who were not given access to the letters. By comparing the scores given by these separate sets of readers, I can measure the impact of the letters on the score at the individual applicant level.

For the students included in the within-subjects experiment, I obtained access to the letters themselves. Text was extracted from the letters, de-identified, and used to construct a number of "features" capturing the content of the letters. I measure the frequency of mentions of academic, athletic, community, and other topics, as well as the overall tone of the letter. I use these features to quantify differences in the types of letters written on behalf of different students and what the application review process in fact values in a student's letter.

My first question is whether letters written on behalf of underrepresented students are distinctive. My investigation is similar in spirit to that undertaken by Alvero et al. (2021), who show that essay content can be used to predict family income. I show that letters submitted on behalf of underrepresented applicants are indeed different from those submitted for their more advantaged peers. Controlling for the quantifiable elements of applications, I show that a one standard deviation increase in the "URM-ness" of an applicant's letters is associated with a 23 percentage point increase in the likelihood that the applicant is indeed from an URM group. For comparison, Alvero et al. (2020) find that the content of application essays can achieve 65% accuracy in classifying applicants as above or below median family income; Alvero et al. (2021) find that essay content explains about one-third more variation in family income than do SAT scores.

However, the distinctiveness of URM students' letters is largely reflective of the fact that their applications are distinctive in many ways captured by traditional admissions criteria like GPAs and SAT scores. When I construct an index from these criteria to best predict URM status, a one standard deviation increase in this index is associated with a 31 percentage point increase in the probability of a URM student. Moreover, the index of traditional applications measures explains nearly $^1/_3$ of the variation in the letter-based index. To isolate the information in the letters that is new, not merely duplicating what is already available, I construct an alternative index of letter features selected to add incremental predictive power to non-letter features of the application. This has much less predictive power – a one standard deviation increase in this index is associated with only a 5 percentage point increase in the URM share.[4]

Second, I use the letter features, in combination with the reader scores from the experiment, to construct a measure of letter strength. Taking advantage of the fact that applications were read both with and without letters, I define a strong letter as one that yields a higher reader score when the application is read with the letters. I allow the data to speak to the relative importance of content, writing style, identity of the recommender, or other aspects of the letter for generating that score. I show that there is substantial variation in quantifiable aspects of letter strength – letters that I score one standard deviation above average are associated with a 5.7 percentage point increase in the likelihood that the reader who observes the letters will give the applicant a higher score than the reader who did not have the letters. However, the letter strength score values different aspects of the letter than those that are predictive of applicant characteristics; my index of letter strength is only weakly correlated (−0.16) with the letter's score for predicting the applicant's demographic group.

Third, I assess the distribution of letter strength among URM and non-URM applicants. Letters written for URM applicants are weaker, on average, than those written for non-URM applicants. About half of this reflects the fact that students who were rated as weaker by readers without access to letters tend to get weaker letters, and the URM applicant pool is weaker on non-letter dimensions than the non-URM pool. Even when I adjust flexibly for the strength of the application without letters, however, I find that URM students' letters are slightly

[2] Subsequent to the period I study, the University of California Regents voted to eliminate the SAT and ACT from UC admissions, though they were in use 2016-17.

[3] Following the pilot study, though before any results from it were available, the UC Regents voted to allow campuses to request letters of recommendation from up to 15% of their applicants each year.

[4] Alvero et al. (2021) also find that essay content is strongly related to SAT scores, but do not measure the incremental information in essays above what is otherwise available.

weaker, on average. The gap in letter strength predicts a 1.3 percentage point gap in the likelihood that the reader who sees the letters will assign a higher score than one who does not.

Finally, I investigate how the inclusion of letters affects application outcomes, and how this varies with URM status. Readers with access to the letters assign somewhat higher scores to URMs relative to non-URMs than would be predicted based on either non-letter application characteristics or the initial read scores. However, this URM benefit is lessened for applicants with the strongest letters, and increased for those with weaker letters. In other words, application readers seem to identify features in the letters that raise their evaluations for the average URM applicant, but within the URM pool they reward strong letters less than they do within the non-URM pool. One plausible interpretation is that the letters reveal information both about applicant strength and about applicant disadvantage; the readers reward both, but put less weight on the strength information for students who come from more disadvantaged backgrounds. I find evidence consistent with this, in that the URM-distinctiveness of the letters is associated with better application outcomes for URM applicants, but even controlling for this the strength of the letters has less effect on URM students' outcomes.

My study relates to the long literature and policy debate regarding the use of quantitative measures in selective college admissions (see, e.g., Bastedo et al., 2018; Black, Denning, & Rothstein, 2021; Bleemer, 2021; Bowen & Bok, 1996; Hout, 2005; Karabel, 2005; Kurlaender et al., 2020; Lemann, 1995; Rothstein, 2004). It also relates to similar debates in other arenas, where policymakers must decide when and how to use subjective information in making consequential decisions. Employers decide who to hire, police decide who to arrest, judges and juries decide whether to convict, and real estate agents decide which houses to show, all based on human judgments, informed by a combination of measurable and unmeasurable factors. Concerns about the possibility that these judgments may introduce bias have been raised in discussions of hiring (Bendick & Nunes 2012; Bertrand & Mullainathan 2004; Biernat & Fuegen 2001; Madera et al., 2009; Schmader et al., 2007), police-citizen interactions (Correll et al., 2007; Gelman et al., 2007; Lai & Zhao 2010; Oliveira & Murphy 2015), and criminal sentencing (Glaser et al., 2015; Rachlinski et al., 2009; Weinberg & Nielsen 2012).

Understanding the role of bias in subjective decisions is difficult, as any given decision could reflect either bias or legitimate use of not easily quantifiable components of the information on which the decision maker relies. The police officer making an arrest decision may interpret the same information differently when it relates to a black or a white suspect, or may be seeing something real in his assessment of the suspect's body language, for example, that raises the (fairly assessed) likelihood of guilt.

The UC Berkeley LOR pilot study provides a unique opportunity to shed light on this issue because it allows me to quasi-experimentally vary the amount of information available to the subjective decision-maker. Moreover, it takes place within a context where the explicit goal is to use the qualitative information in an unbiased way, to identify positive aspects of students' applications that may not be apparent from other information. Thus, my results shed light on the potential to use qualitative information to increase fairness of review processes, but can be seen as an upper bound on what might be found in other processes where the goals and training are less clear.

2. The UC Berkeley letters of recommendation pilot

2.1. Admissions at UCB

UCB is one of the most prestigious universities in the country. US News ranks it 22nd among all universities and 2nd among public universities, and measures its undergraduate admissions rate as 17% (US News 2021). US News also reports that the 25th and 75th percentile SAT scores among admitted students were 1310 and 1530, respectively, corresponding to the 87th and 99th percentiles of all SAT-takers (Board, 2020). This is comparable to NYU or Rochester. UCB is one of the largest of the highly selective universities, with over 30,000 undergraduates.

The university's size, along with budget constraints, means that the admissions process is more routinized than at smaller private colleges and universities. Historically, admissions were largely mechanical – Bleemer, 2022 documents that in the mid-1990s, at least half of Berkeley admissions were based on SAT and high school GPAs alone. More recently, Berkeley has adopted "holistic review" (Hout, 2005), and each application is read and scored by two separate readers. However, even this holistic review is an industrial process. The university hires dozens of temporary application readers each spring, and anticipates that they will spend only a few minutes on each application. Where Stevens (2007) assigns a large role at the small liberal arts college he studies to "committee," a group discussion of individual applicants that leads to a consensus decision, there is no such consensus-building process at UCB. Rather, the two separate reader scores are mechanically aggregated into an admissions decision, without a role for group deliberation.

Because the context for the LOR pilot is important to understanding its effects, I provide some detail about the admissions process. Students use a single application to apply to any of the University of California (UC) campuses, specifying the campuses to which they wish to apply and a college and/or major at each campus. With the application, they submit their transcripts, grade point averages, test scores, and essays written in response to system-wide prompts. The central UC admissions office then forwards applications to each campus that the applicant selected.

At UCB, each application is read twice. The first evaluations are conducted by a large pool of readers that include both permanent UCB admissions staff and outsiders hired on a short-term, piecework basis. Readers score applications on a three-point scale – Yes, Possible, or No. They are instructed to consider the whole application package rather than any particular numerical components in isolation, and work largely independently with occasional "norming sessions" where the team discusses the scores given to a small number of applications with the goal of establishing common scales. Applications are then evaluated a second time, this time by a more limited pool of more experienced reviewers. These readers apply the same holistic review, but this time have access to the first readers' scores and any comments in addition to all other application materials.

These two rounds of reading yield nine possible scores (e.g., Yes-Yes, Possible-Yes, etc.). Files then go to the central admissions staff for decisions. These decisions incorporate capacity constraints for the various undergraduate colleges – for example, the College of Engineering is much more competitive than the College of Letters and Science – and for some majors. Within each college or major, decisions usually respect the rankings established by the two reader scores. That is, all applicants to a division with two Yes scores are admitted before any students with a single Yes. Within rankings groups (e.g., among those receiving two "Possible" scores), tiebreaking is done by the admissions staff. Tiebreaking is done en masse, and does not involve careful consideration of each individual application. An implication is that the only way that subjective elements of an application, or any implicit bias of admissions staff, can influence the admissions decision is through the reader scores.

2.2. Underrepresented applicants

A primary motivation of the UCB LOR pilot was to assess whether LORs could be used to identify underrepresented applicants who were strong enough to admit but unlikely to be admitted without the LORs. The UCB admissions office tracks four indicators of applicant disadvantage: Low income families, parents who did not attend college, low-test-score high schools, and underrepresented racial and ethnic groups. Three of these four are included on the application materials made available to readers. Race and ethnicity are not included, as the California constitution prohibits the consideration of either in admissions or hiring, though in some cases readers may be able to infer them from

names and other applicant characteristics. For most of this paper, I will group the four categories together into a "URM" composite, consisting of any student who falls in any of the groups. Overall, URMs comprise about two-fifths of in-state, non-athlete applicants, with about one-fifth in each of the four overlapping categories.

2.3. The LOR pilot

The pilot study that I examine was implemented in 2016–2017, for applications submitted in November 2016 for matriculation in Fall 2017. Hereafter, I refer to this as the "2017″ admissions cycle. The 2017 study followed a smaller pilot in the previous cycle. Out of concerns that requests for LORs would be burdensome for applicants' teachers and counselors, not all applicants were asked to submit letters. Instead, students applied without letters in November, and in early December a subset of applicants were invited to provide letters, which if submitted were made available to the second-round readers.

The goal was to invite only students who were marginal for admission, but of course it is difficult to identify marginal students before applications are reviewed. A two-pronged approach was used for quickly selecting students to invite. First, the first round of application reviews was accelerated to generate scores for as many applicants as possible by early December. Any students who were scored as "Possible" by this point were invited to submit letters. In addition, the admissions office fit a statistical model to the previous year's scores and used this to predict students with high probabilities, based on their quantifiable application measures, to receive "Possible" scores.[5] All students selected by the model were invited to submit letters, regardless of the timing or outcome of the first reader's evaluation. In total, 30% of (in-state, non-athlete) applicants were invited to submit letters.

Applicants invited to submit letters were assured that a failure to do so would not be counted against them, and were asked to provide names and contact information of letter writers, who were then asked directly for their letters by the application system. 79% of invited students provided names of recommenders, and 77% (97.5% of those who provided names) wound up with at least one submitted letter.[6] Rothstein (2017) shows that underrepresented applicants were much less likely to request letters, but that this was concentrated among students who were unlikely to be admitted in any case.

The first round of application review was completed before letters arrived. The second round considered letters when they were available. Readers used the same three-point scale and were instructed not to hold a student's letters against them and not to penalize students without letters.

2.4. The within-subjects experiment

As part of the pilot, 10,000 students who had returned letters – about half of the total – were selected for inclusion in a more detailed study. I hereafter refer to these students as "the study sample." Their applications were reviewed a third time for the study, with these third reader scores playing no role in actual admissions decisions. The third round was conducted by readers recruited from the group who conducted the second reads, after decisions had been made but otherwise under conditions designed to replicate as closely as possible the second reads as they would have occurred had LORs not been available.[7]

Like the first and second readers, the study readers (the "third readers") scored applications on the Yes/Possible/No scale. Their scores were intended to represent the counterfactual outcome that the student would have obtained had LORs not been considered. The "within subjects" comparison between the actual second read scores and the third reader scores from this simulation thus provides an estimate of the effect of LORs at the level of the individual student. As it turned out, the third readers were less generous in their scoring than either the first or the second readers. Overall, 26% of applicants in the study sample received higher scores from the 2nd readers and 15% received higher scores from the 3rd readers; 59% received the same score from each.[8] These student-level treatment effects are quite noisy, however, as an application's second and third reader were different people who may have prioritized different aspects of the applications. Moreover, while the lower score distribution in the 3rd round may indicate that the 2nd readers were positively influenced by the letter content, it is also possible that the score norms that readers actually used changed a bit between rounds. Thus, I do not examine student-level treatment effects, but use them to estimate the average effect of letters for students with particular characteristics (e.g., test scores, or letter content).

Table 1 presents summary statistics for the full sample of applicants and for three progressively narrower subgroups: Those invited to submit letters, those who actually submitted them, and the study sample. Invited students had slightly higher GPAs but lower SATs than the average applicant, and were substantially more likely to come from the URM groups (55% vs 42%). They were also much less likely to have received "No" scores and somewhat less likely to have received "Yes" scores from the first readers – as expected, given the role of a "Possible" score in selecting students for the LOR pilot. Those who actually submitted letters were positively selected on both GPAs and SATs relative to those invited, and were less likely to be URM (though the URM share was still higher than in the applicant pool as a whole). Last, as noted, in the experimental sample the third readers' scores, without LORs, were a bit lower than the second readers' scores, with LORs, for the same students.

2.5. The admissibility index

Even with holistic admissions, quantifiable information plays a large role. It is possible for a student with very weak high school GPA, test scores, and course-taking, but strong essays or LORs, to be admitted, but it is quite unlikely. It is useful for my purposes to have a unidimensional summary of the strength of the quantifiable portions of the application. I

[5] This model was a logistic regression where the outcome was an indicator for a "Possible" score, with both "Yes" and "No" scores coded as failures. Explanatory variables consisted of quantifiable admissions measures (test scores, grades) and demographic characteristics (school composition), but not individual race and gender. I used the same variable list to construct the admissibility index discussed below. Applicants with predicted probabilities of "possible" scores above 50% were invited to submit letters.

[6] Many of the students who did not provide names of recommenders withdrew their applications, perhaps due to an early admission decision elsewhere. For a student already applying to other UC campuses, adding an application to Berkeley involved only checking an additional box on the application form and paying an application fee (which could be waived). There appear to have been many students who applied to Berkeley but were not serious about these applications, and who did not submit letters.

[7] Readers participating in the study were compensated at a slightly higher piece rate than in the regular cycle. A lead reader led norming sessions, as in the regular process. Applications were not assigned to third readers who had previously reviewed the application in either the first or second rounds; a handful of the 10,000 students in the study sample are excluded from my analysis sample due to inadvertent violations of this rule.

[8] For comparison, if second and third read scores had been randomly and independently assigned to applicants using the marginal distributions seen in the sample, 38% of students would have received the same score in both rounds. I am not aware of an estimate of the inter-rater reliability of the UCB read scores holding the available information constant, but one way to estimate it is to compare the scores given to students in the study sample by the first and third readers, neither of whom had access to the letters. 65% of students received the same scores in these two rounds. Note, however, that the 3rd readers, like the 2nd readers, had access to the 1st readers' scores, which may have nudged them toward a higher rate of concordance than would occur with fully independent assessments.

Table 1
Summary statistics.

	All applicants	Invited to submit letters	Submitted letters	Experimental sample
	(1)	(2)	(3)	(4)
N	85,061	25,121	19,295	9,993
College				
Chemistry	4%	4%	4%	3%
Engineering	23%	22%	21%	19%
Environmental Design	2%	2%	2%	2%
Letters & Sciences	62%	63%	65%	67%
Natural Resources	9%	9%	9%	10%
Underrepresented students				
Any underrepresented category	42%	55%	49%	62%
Black or Hispanic	24%	35%	30%	40%
First generation	16%	28%	23%	32%
Application fee waiver	22%	36%	30%	40%
Low-ranking high school	26%	46%	39%	39%
Application characteristics				
GPA	3.68	3.74	3.79	3.75
SD	[0.35]	[0.31]	[0.25]	[0.28]
SAT composite	1925	1897	1944	1883
SD	[261]	[297]	[269]	[282]
Admissions index	16%	20%	22%	22%
SD	[22%]	[22%]	[22%]	[22%]
Letters of recommendation				
Invited	30%	100%	100%	100%
Requested	23%	79%	100%	100%
Submitted letters	23%	77%	100%	100%
Reader scores				
Reader 1 (without letters)				
No	62%	25%	17%	24%
Possible	24%	72%	79%	71%
Yes	14%	3%	4%	5%
Reader 2 (with letters)				
No	63%	40%	32%	37%
Possible	21%	43%	48%	44%
Yes	15%	17%	20%	19%
Reader 3 (without letters)				
No				42%
Possible				45%
Yes				13%

Notes: All statistics pertain to in-state applicants who are not recruited athletes.

used data from the 2016 admissions cycle to construct this. I estimated a logistic regression taking admission as the dependent variable and considering as explanatory factors the SAT score, high school GPA, average SAT scores at the applicant's high school, the applicant's course-taking relative to what was available in the high school, the high school's test score rank (based on average scores on state accountability tests), the average parental income at the high school, an indicator for a high school at which fewer than 5% of graduates apply to the UC, and the applicant's parents' education and income. All of these are measures that are featured prominently in the information provided to readers for use in scoring applications. The model also includes fixed effects for the college and major applied to. I use this logistic model to generate a predicted probability of admission for each applicant in the 2017 cycle, and I refer to this as the "Admissions Index," or AI.

The AI is scaled as a predicted admissions probability. In the overall pool of applicants to UCB, the average AI is 0.16 and the standard deviation is 0.22. A large share of applicants has very low AIs – 44% of applicants have AIs below 0.05, while only 27% have AIs above 0.2. The former applicants are very unlikely to be admitted, and are also unlikely to receive the positive initial evaluation needed to support an LOR invitation. It is important to emphasize that these strong relationships do not indicate a failure of the holistic review process. Rather, the AI simply summarizes the outcomes of that process, and captures the quantifiable characteristics of students who typically did well or poorly in holistic review. The variation in outcomes among students with the same AIs (for all AIs greater than 0 and less than 1) reflects other factors that are captured by holistic review but not by the AI.

On average, URM applicants have weaker academic qualifications than do non-URM applicants. Fig. 1 shows AI distributions for the two groups, both for the full sample and the experimental sample. The mean and median AI among all non-URM applicants are 20% and 9%, respectively, while among URM applicants the mean is 12% and the median is 5%. The AI distribution is higher in the experimental sample, selected on the basis of the initial application evaluation discussed above, due largely to reduction in the density at the left tail. Only 27% of experimental sample applicants have AIs below 0.05, much lower than the 44% in the full sample. This reduction is made up by higher density in the upper portion of the distribution, around 0.25: 40% of the experimental sample has AIs above 0.2. Gaps between URM and non-URM applicants are if anything larger in the experimental sample: The median AI is 21% for non-URM applicants and 10% for URM applicants.

2.6. *Letter content*

Following the pilot study, the admissions office provided to me the text from the letters that were received for each of the students in the study sample, extracted from the original PDFs, along with applicants' responses about the identity of the letter writers (e.g., the writer's relationship to the applicant). To preserve applicant confidentiality, the letter texts were de-identified, using a procedure described in detail in the appendix. Most applicants had two letters, so after dropping students whose letters could not be parsed I am left with a corpus of 17,645 letters written on behalf of 9841 applicants.

I use a "bag of words" approach to assign features to each letter, using both off-the-shelf dictionaries and some developed specifically for this study. The off-the-shelf dictionaries come from the Linguistic Inquiry and Word Count (LIWC) package, which has been used for several previous studies of LORs (e.g., Houser & Lemmons, 2018; Madera et al., 2009; Schmader et al., 2007) and college essays (Alvero et al., 2021). The package counts the number of words from each dictionary that appear in each letter, accounting for the use of word stems. I use 23 features from this package. These come in several types. One set of dictionaries focuses on specific topics, such as "achievement," "drives," and "power." Another focuses on the writing itself, with measures of the number of words longer than six letters, the number of words per sentence, and the length of the letter. A third attempts to summarize the tone of the letter, with measures of "analytical thinking" and "emotional tone," and of the degree to which the language is present-, past-, or future-focused. Pennebaker et al (2015) describes the development and validation of the LIWC dictionaries.

I also developed some dictionaries specifically for this project, with the goal of capturing domains that may be important in college application LORs but not as important in the more general texts used to develop LIWC. I began by generating counts of 300 of the most common words or word stems in the LOR corpus, plus a list of 98 hand-curated adjectives. Thus, separate variables measure the number of uses of the words "volunteer," "conscientous," "ability", and "challenge." Next, I and several research assistants constructed ten dictionaries corresponding to topical areas more important in LORs than in the general writing that LIWC is designed to measure, such as academics, community, sports, and humanities. For example, my "STEM" dictionary includes words like physics, computer, statistic, and laboratory; the "grit" dictionary includes drive, enthusiasm, train, focus, effort, etc. Finally, I also construct measures of the letter writer's relationship to the student,

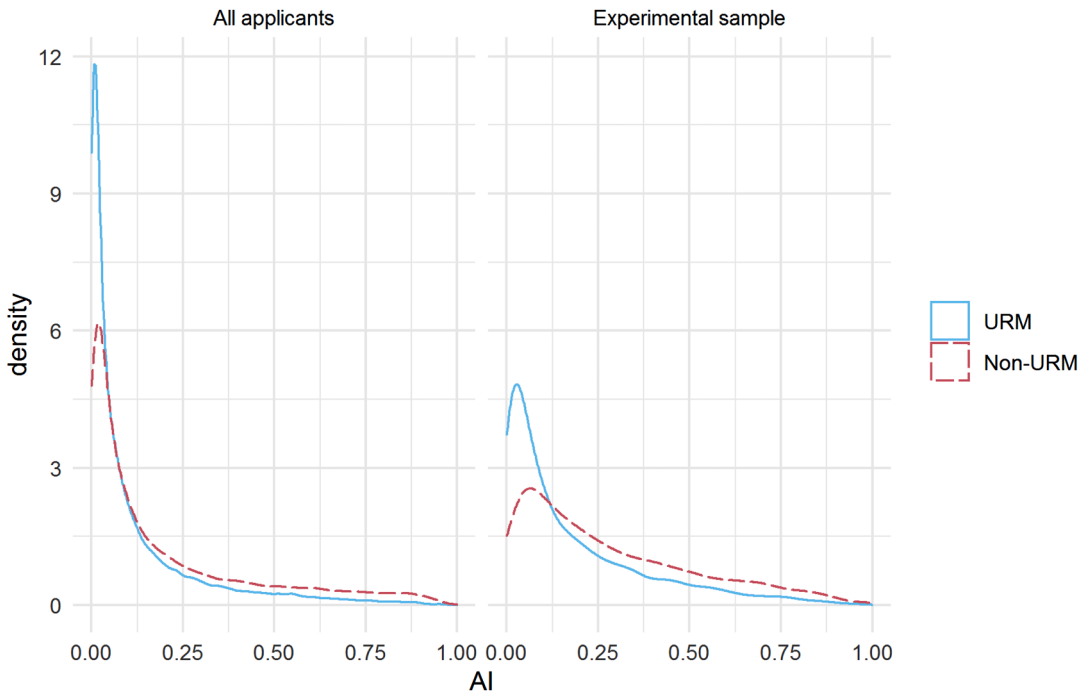

Fig. 1. Distribution of admissions index, by sample and applicant URM status
Notes: Admissions index is the predicted probability of admission given applicant characteristics, using fitted model from previous admissions year.

as reported by the student: Teacher, counselor, coach, employer, or other.

The Appendix includes full lists of the words included in each of the ten bespoke dictionaries; see the LIWC documentation (Pennebaker et al., 2015) for its word lists. Table 2 presents summary statistics for the bespoke word counts. The average student's letters include 31 uses of "academic" words, 12 of "extracurricular" words, and 19 of "grit" words. URM students' letters include fewer uses of each category of words except for those relating to family.

Table 2 also shows a principal components analysis of these features along with a few key application features. The first component, accounting for 26% of the total variance, seems to be a broad measure of academic strength, with substantial weight on GPAs, SATs, and counts of academic-related words. The second component, 16% of total variance, seems to distinguish the purely quantitative aspects of the application from the letters, with positive weight on non-letter characteristics and negative weight on all of the letter features. The third component, 7% of the variation, seems to distinguish applicants whose letters emphasize humanities and language from those whose letters emphasize sports and/or STEM topics.

3. Are letters written for underrepresented students different?

The first question I investigate is whether there are measurable differences between the letters written for URM and non-URM students. There are several reasons to anticipate such differences, even before seeing the summary statistics in Table 2. First, as hypothesized by the supporters of the LOR policy change, some students will have overcome challenges on the way to college that their recommenders find worthy of mention, and this may be more common among URM than non-URM applicants. Second, as suggested by the opponents of the policy, the recommenders for URM students may be more over-stretched, have weaker writing skills, and be less familiar with the selective college application process; all of these might translate into differences in the letters they produce. Third, and perhaps most important, URM and non-URM applicants are different in many ways, including in ways relevant to the admissions decision, as evidenced by the differences in AI distributions seen in Fig. 1 and the gaps in test scores and GPAs in Table 2. If we think that one role of the letter is to express in words the academic accomplishments indicated by the numerical metrics, we might expect the letters written on behalf of URM students to be less effusive about academic achievements, on average, than those written for non-URM students, even when backgrounds and letter writers are the same.

I treat the third source of differences as a confounder in the analysis, as LORs are never considered in admissions on their own, but only as complements to the quantitative elements of the application.[9] Thus, my question is whether letters written for URM students differ from those written for non-URM students *with similar quantitative elements of the application portfolio*. To measure this, I ask whether letter features predict a student's URM status, over and above the information contained in the rest of the application.

Specifically, let X_i represent the quantifiable components of the application of student i, let W_i be a vector of letter features, and let URM_i be an indicator for a URM applicant. I fit three prediction models. The first model uses just the letter features to predict URM. I use this model to generate a predictive probability that a student with letter features W comes from a URM group,

$$D_i^W \equiv \widehat{P}[URM_i \mid W_i]$$

D^W(for "distinctive) measures the degree to which the letter content identifies the applicant as likely to be URM (D^W close to one) or not (D^W close to zero).

The second model is the same, but uses only non-letter features X to predict URM. It generates a distinctiveness measure D^X:

$$D_i^X \equiv \widehat{P}[URM_i \mid X_i].$$

Finally, a third model uses both X and W:

[9] This is a contrast to Alvero et al. (2020, 2021), who do not adjust their analysis of essay content for information that overlaps with quantitative application measures.

Table 2
Summary statistics for letter features, experimental sample.

	Mean	SD	Non URM - URM difference	Correlation with AI	Principal component analysis		
					Comp. 1	Comp. 2	Comp. 3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
GPA (unweighted)	3.75	0.28	0.15	0.55	0.29	0.14	−0.05
GPA (weighted)	4.18	0.36	0.23	0.63	0.35	0.19	−0.06
Honors classes	17.6	6.6	3.5	0.50	0.28	0.16	−0.04
SAT Math	651	106	119	0.50	0.37	0.21	−0.03
SAT Reading	638	109	114	0.53	0.35	0.20	0.07
SAT Writing	596	104	88	0.60	0.31	0.17	0.09
ACT composite	28.2	4.4	4.2	0.36	0.30	0.19	0.08
Letter topics (word counts)							
Academic	31.1	15.0	4.2	0.20	0.25	−0.35	0.07
Arts	7.0	5.8	2.0	0.10	0.18	−0.26	0.14
Extracurriculars	12.4	8.3	0.8	0.08	0.18	−0.38	−0.18
Family	12.1	7.7	−0.4	0.02	0.13	−0.42	0.03
Grit	19.0	9.5	1.6	0.07	0.19	−0.39	−0.12
Humanities	3.2	3.9	0.8	0.09	0.12	−0.14	0.60
Language	0.6	1.7	0.3	0.03	0.07	−0.08	0.42
Social Sciences	0.8	1.6	0.1	−0.02	0.06	−0.16	0.26
Sports	4.8	6.0	1.3	0.04	0.13	−0.22	−0.37
STEM	6.7	7.7	3.1	0.22	0.20	−0.07	−0.39
PCA share of explained variance					26%	16%	7%

Notes: Students typically submit only one of the SAT and ACT. Students not submitting one of the tests are assigned the mean value among those who do.

$D_i^{X,W} \equiv \widehat{P}[URM_i \mid X_i,\ W_i]$.

The degree to which the letters themselves are distinctive to students of a particular group, over and above what one would predict based solely on the quantifiable application features that are available to readers even in the absence of letters, is captured by the difference between the second and third predictions, $\Delta_i^W \equiv D_i^{X,W} - D_i^X$. A letter that is distinctive of a type that is written for URM students not otherwise identifiable by their *X*s will lead to a Δ^W substantially above zero, while one that is more typical of non-URM students will lead to a negative Δ^W.

To fit the three prediction models, I use random forests (James et al., 2013). Random forests allow for arbitrary non-linearity and interactions in the relationships between the predictor variables and the variable being predicted, and use variation across trees given access to different subsets of predictors to control over-fitting. Nevertheless, even random forests can over-fit to chance features of the sample, which can bias upward the predictive power of the features for the outcome examined. To guard against this, I divide the study sample into two random subsamples. I use one, consisting of 75% of the original sample and labeled the "training" sample, to fit the random forest models. I then use the remaining observations, the "test" sample, to assess the accuracy of the predictions. I find that the X features alone can predict 40.1% of the variation of URM in the training sample, and the index is quite reliable out of sample, explaining 39.5% of the variation in the test sample. Adding W to the model increases the explained share of variation only to 41.0% in the training sample and 40.5% in the test sample. Thus, the inclusion of letter features raises the explained share of variance by only about one percentage point, indicating a modest degree of distinctiveness of the letters written for URM candidates.

Fig. 2 shows the distribution of Δ^W among URM and non-URM candidates. These are only modestly different. The average URM candidate in the test sample has a predicted URM probability, based solely on X, of $D^X = 0.759$. His or her letters raise that predicted probability that the applicant is URM by only Δ^W=0.004 percentage points. The average non-URM candidate, by contrast, has $D^X = 0.364$, while the letters reduce the predicted probability by Δ^W=−0.006. There is meaningful dispersion around this mean, however. The standard deviation of Δ^W is 0.06.

Fig. 2 also includes a series that reweights the non-URM candidates' AI distribution to more closely represent the URM students, to ensure that we are comparing candidates who are the same but for their letters. This makes little difference.

One reason that letters may have little incremental predictive power is that URM status is already well predicted by the X features. The second panel of Fig. 2 shows the distribution of Δ^W in the subsample of students who were not already identified as almost certainly URM by their X features (i.e., those who have $D^X < 0.9$). Again, we see substantial dispersion, but a relatively small difference between the distributions for URM and non-URM students.

Table 3 presents several linear probability models that use D^W, D^X, and Δ^W to predict URM, all estimated in the test sample. Column 1 shows a model that uses D^W as the sole predictor. This has a coefficient of 1.62, indicating that a 0.01 increase in distinctiveness of the letter features is associated with a 1.62 percentage point increase in the likelihood of the candidate being a URM.[10] This explains 22 percent of the variation in URM status; the average URM has a predicted probability from this model of being URM of 69%, as compared with 47% for the average non-URM applicant. Thus, letters for URM applicants are indeed quite distinctive from those written for non-URM applicants.

Column 2 uses just the index constructed from non-letter features, D^X. This explains nearly twice as much of the variation in URM status: The average URM applicant has a predicted probability of 76%, as compared with 36% for non-URMs. Column 3 adds Δ^W to the model. This has a positive and statistically significant coefficient – letter features do improve predictions of URM status. However, it increases the R^2 by only 0.01 and only incrementally improves the ability to distinguish between URM and non-URM applicants. Columns 4 and 5 add the AI as a control. While it is a statistically significant predictor – controlling for D^X and Δ^W, the AI is positively associated with being a URM – it does not change the overall patterns and adds only minimally to the ability to distinguish the two groups of applicants. Finally, columns 6 and 7 include Δ^W without D^X, first without and then with the AI control. The letter-based index is a statistically significant predictor of URM status, but it does not vary enough to meaningfully distinguish the two groups of applicants. Without the D^X control, the AI is negatively correlated

[10] In principle, in the training sample the coefficient of this regression should be 1: A 0.01 increase in the predicted likelihood of being URM should be associated with a one percentage point increase in the actual likelihood. In fact, the coefficient is 1.58 in the training sample. This reflects a common occurrence, which is that random forest models are not always well calibrated. Thus, the units of D^W need to be rescaled by approximately 1.6 to be interpreted as predicted probabilities.

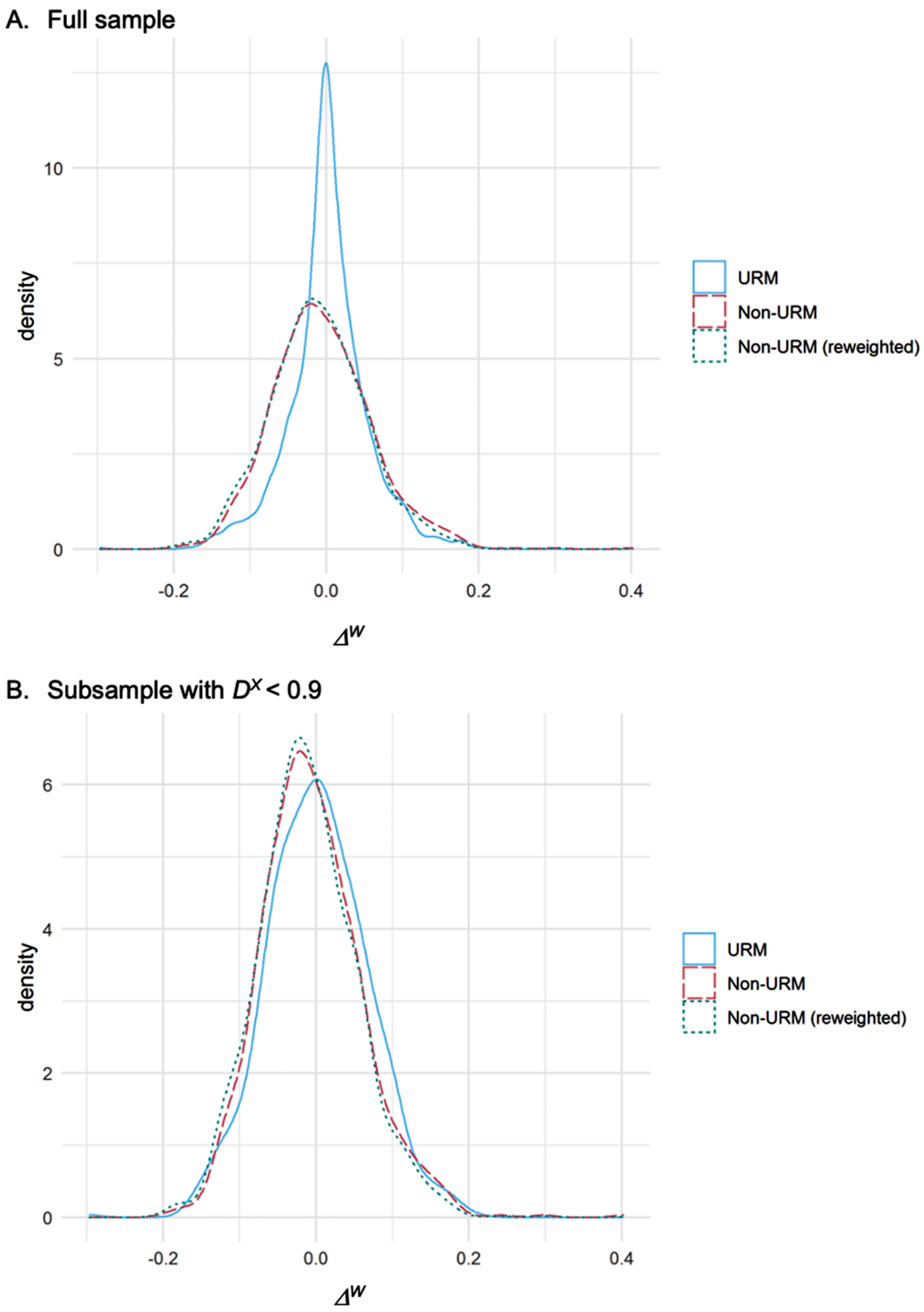

Fig. 2. Distribution of URM distinctiveness of letters (Δ^W), by URM status, experimental sample

Notes: Figure shows the density of the letter distinctiveness index, the increment to the predicted probability that a student is URM from considering the letter features. The reweighted series reweights the non-URM students to have a similar Admissions Index (AI) distribution as the URM students. In panel B, the sample is limited to students for whom the predicted URM probability given non-letter features (D^X) is less than 0.9.

with URM status, as expected from Fig. 1.

Overall, Table 3 indicates that there is indeed information in letters that can distinguish URM from non-URM candidates. However, there is not a great deal of it: The Δ^W index on its own explains only 0.4% of the variance of a URM indicator, and its incremental explanatory power once D^X is controlled is just 1%.

Table 4 presents models that further illustrate the magnitude of differences in Δ^W between URM and non-URM applicants. To estimate these, I regress Δ^W on URM, again adding controls for the AI and then for D^X. The difference in the letter URM score Δ^W between URM and non-URM applicants is only 0.007, on average. This more than doubles when I add controls, but remains under two percentage points. Thus, while letter distinctiveness can be reliably identified, it is not very dramatic – readers can already identify URM candidates with a high degree of accuracy, and distinctive letters add only a little to this.

Thus far, I have grouped the four URM groups together. I can instead treat them separately, estimating separate prediction models for students with low-education parents, for students from low test score high schools, and so on. The resulting Δ^W are modestly correlated with each other, with correlations ranging from 0.27 to 0.43. Tables 5 and 6 present estimates corresponding to those in Tables 3 and 4 for each of the separate Δ^W indexes. Letters written for students from low-test-score

Table 3
Predicting URM status using application and letter features, in test sample.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Letter-based index of URM, unadjusted (D^W)	1.62						
	(0.06)						
Application-based index of URM (D^X)		1.00	1.01	1.09	1.10		
		(0.02)	(0.02)	(0.03)	(0.03)		
Letter-based index of URM, adjusted for application features (Δ^W)			0.87		0.83	0.52	0.65
			(0.13)		(0.13)	(0.17)	(0.17)
Academic Index				0.28	0.27		−0.48
				(0.04)	(0.04)		(0.04)
R^2	0.223	0.395	0.405	0.406	0.416	0.004	0.048
Mean predicted probability of URM							
Non-URM students	0.468	0.364	0.358	0.357	0.352	0.600	0.573
URM students	0.691	0.759	0.763	0.764	0.768	0.604	0.621

Notes: All columns report linear probability models where the dependent variable is an indicator for a URM student. The sample is the "test" sample, held out from the construction of the letter-based indices. $N = 2{,}460$.

Table 4
Predicting letter URM-ness index.

	(1)	(2)	(3)
URM	0.007	0.009	0.019
	(0.002)	(0.002)	(0.003)
Academic Index		0.023	0.008
		(0.006)	(0.006)
Application-based index of URM (D^X)			−0.029
			(0.005)
R^2	0.004	0.011	0.022

Notes: Dependent variable in all columns is the increment to predicted applicant URM status from including letter features in the prediction, Δ^W. The sample is the "test" sample, held out from the prediction model. $N = 2{,}460$.

high schools are more distinctive from those written for students from other high schools than are those for the other groups. There is little information in letters, however, about parental education, family income, or student racial or ethnic identity.

This distinctiveness of letters from low-performing high schools is perhaps consistent with the hypothesis that teachers and counselors at these schools are not positioned to be good letter writers. However, thus far I have measured only letter distinctiveness, not strength – the letters could be distinctive because they effectively highlight how the student's ability to overcome obstacles prepares them for college. I turn next to an analysis of letter strength.

4. Measuring letter strength as interpreted by application readers

I have established thus far that letters written on behalf of URM applicants differ systematically, though not majorly, from those written on behalf of non-URM applicants. But different need not be better or worse; readers can infer an applicant's URM status with high accuracy even without the letters, may not consider the "URM-ness" of an applicant's letters in scoring his or her application, and may (or may not) be able to extract substantial information from the letters that is unrelated to the applicant's URM status.

The next step of the analysis is to measure the features of letters that lead to stronger and weaker reader scores. As in the above analysis of the distinctiveness of letters written for URM students, a first-order challenge is to disentangle the effects of letter features from effects of the rest of the application that might be correlated with those letter features. For this analysis, I can take advantage of the within-subjects experiment. Rather than predicting a reader score, where it would be necessary to disentangle the effects of letters from that of other features, I instead examine the difference between the score given in the second round, when letters were available, and that given in the experimental third round, when they were not. I ask which letter features predict a stronger score from the second reader. That is, I fit another random forest model, this time using just letter features as predictors (i.e., no Xs), where the objective is to predict whether the applicant received a higher score from R2 than from R3.[11] I exclude from this analysis those applicants who received "Yes" scores from R3, as there was nothing the letters could have contributed to raising these students' scores.

Table 5
Predicting URM subgroups using application and letter features.

	(1)	(2)	(3)
A. Low API high school			
Letter-based index of low API, adjusted for application features (Δ^W)	0.86	0.84	0.95
	(0.16)	(0.16)	(0.13)
Academic Index		−0.28	0.31
		(0.04)	(0.04)
Application-based index of low API (D^X)			1.10
			(0.03)
R^2	0.012	0.028	0.350
B. Fee waiver			
Letter-based index of fee waiver, adjusted for application features (Δ^W)	0.54	0.66	0.85
	(0.15)	(0.15)	(0.12)
Academic Index		−0.48	0.23
		(0.05)	(0.04)
Application-based index of fee waiver (D^X)			1.09
			(0.03)
R^2	0.005	0.049	0.397
C. First generation			
Letter-based index of first generation, adjusted for application features (Δ^W)	0.27	0.37	1.17
	(0.15)	(0.15)	(0.13)
Academic Index		−0.38	0.22
		(0.04)	(0.04)
Application-based index of first generation (D^X)			1.15
			(0.03)
R^2	0.001	0.032	0.334
D. Underrepresented racial or ethnic group			
Letter-based index of racial/ethnic minority, adjusted for application features (Δ^W)	0.22	0.35	0.58
	(0.17)	(0.17)	(0.13)
Academic Index		−0.58	0.11
		(0.04)	(0.04)
Application-based index of racial/ethnic minority (D^X)			1.08
			(0.03)
R^2	0.001	0.066	0.395

Notes: Samples and specifications are as in Table 3. Dependent variables in each panel are indicators for the indicated student characteristic. $N = 2{,}460$.

[11] Recall that random forest models by construction allow for arbitrary interactions among the included variables. Thus, if readers value "well-rounded" applicants and thus reward letters that include discussions of both athletic and academic interests, the prediction model should in principle capture that.

Table 6
Predicting distinctive letter scores.

	(1)	(2)	(3)
A. Low API high school			
Low API high school	0.014	0.014	0.023
	(0.003)	(0.003)	(0.003)
Academic Index		0.000	−0.014
		(0.006)	(0.006)
Application-based index of low API (D^X)			−0.031
			(0.006)
R^2	0.012	0.012	0.022
B. Fee waiver			
Fee waiver applicant	0.009	0.012	0.024
	(0.003)	(0.003)	(0.003)
Academic Index		0.029	0.011
		(0.006)	(0.007)
Application-based index of fee waiver (D^X)			−0.036
			(0.006)
R^2	0.005	0.014	0.028
C. First generation			
First generation applicant	0.005	0.007	0.028
	(0.003)	(0.003)	(0.003)
Academic Index		0.024	−0.011
		(0.006)	(0.006)
Application-based index of first generation (D^X)			−0.081
			(0.006)
R^2	0.001	0.008	0.072
D. Underrepresented racial or ethnic group			
Racial/ethnic minority	0.003	0.005	0.013
	(0.002)	(0.002)	(0.003)
Academic Index		0.019	0.008
		(0.006)	(0.006)
Application-based index of racial/ethnic minority (D^X)			−0.024
			(0.005)
R^2	0.001	0.005	0.013

Notes: Sample and specifications are as in Table 4. Dependent variable in each panel is the increment to the predicted probability that an applicant has the indicated identity from including letter features in the prediction.

As in Section III, I use a split-sample strategy to manage over-fitting. I estimate the prediction model using the training sample, then use the test sample to assess the quality of the predictions. The predicted values from the model are

$$\Sigma_i \equiv \widehat{P}[1(R2_i > R3_i) \mid W_i;\ R3_i < Yes]$$

Σ_i can be interpreted as an index of letter strength: The higher is Σ_i, the more likely is a student with letter features W_i to benefit from the inclusion of letters in the admissions process. Although the model is fit only to applicants who do not receive "Yes" scores from R3, I use it to generate measures of letter strength for all applicants.

Table 7 presents regressions of various admissions outcomes on Σ_i. Of note is the model for the outcome of a higher score from R2 than from R3, the same measure used to train the Σ_i index. This coefficient is smaller than 1 in the training sample; random forest models, unlike OLS, do not generate unbiased predictions. This indicates a modest amount of overfitting, but also substantial signal value in the letter strength index. I construct a new index $\widetilde{\Sigma}_i = 0.8 * \Sigma_i$ that scales down the original index by this coefficient. This index measures the strength of the applicant's letters in units of the probability that the letters predict that the reader score will rise when they are considered. The letter strength index is also strongly positively correlated with the second reader's score on its own, with the simple difference between the 2nd and 3rd readers' scores (converting them each to a 1–2–3 scale), and with the eventual admissions outcome.

Fig. 3 shows the distribution of $\widetilde{\Sigma}_i$ in the full sample, for both URM and non-URM students. There is substantial variation in letter strength – a 25th percentile letter is associated with a 25 percent chance of getting a stronger score from reader 2 than from reader 3, while a student with a 75th percentile letter has a 33 percent chance of this. The standard deviation of $\widetilde{\Sigma}_i$ is 0.057.

I can also examine the specific letter features that seem to be important to measured letter strength. Pure length is an important factor, with longer letters tending to be stronger. Features measuring a focus of the letters on academic strength – counts of words like "AP," "STEM," "academic," "research," "science," and "intellectual" – seem to be strongly represented among the features with high importance. Letter writer identity is comparatively unimportant – none of the variables measuring the student's relationship to the letter writer appear high on the list.

The correlation between the letter strength score $\widetilde{\Sigma}_i$ and the distinctiveness score Δ_i^W is a fairly weak −0.16.

5. Are letters written for underrepresented students stronger or weaker than those written for other students?

Fig. 3 shows that there are notable differences in the strength of letters written on behalf of URM and non-URM applicants. The average non-URM applicant has a letter strength score $\widetilde{\Sigma}_i$ of 0.303, while the average URM applicant has only 0.278. The non-URM mean falls only to 0.300 when this group is reweighted to the URM AI distribution.

Table 8 presents regressions that investigate this. The raw gap in $\widetilde{\Sigma}_i$ between URM and non-URM applicants is 0.026, indicating a 2.6 percentage point difference in the probability of receiving a higher score from reader 2 than from reader 3. This is nearly half of a standard deviation of $\widetilde{\Sigma}_i$, a substantial difference.

Columns 2–4 add controls meant to capture the strength of the application without the letters. Column 2 adds the AI, while column 3

Table 7
Letter strength index (Σ_i) and application outcomes.

	Regression on Σ_i				Regression on Σ_i			
	Coeff. (1)	SE (2)	N (3)	R^2 (4)	Coeff. (5)	SE (6)	N (7)	R^2 (8)
A. Training sample								
Read 2 score > Read 3 score	0.744	0.067	6,397	0.019	1.085	0.098	6,397	0.019
Read 2 score (1–3)	2.319	0.098	7,381	0.071	3.382	0.142	7,381	0.071
Read 2 score - Read 3 score	0.532	0.097	7,381	0.004	0.776	0.141	7,381	0.004
Admission	0.903	0.063	7,381	0.027	1.316	0.091	7,381	0.027
B. Test sample								
Read 2 score > Read 3 score	0.686	0.120	2,137	0.015	1.000	0.175	2,137	0.015
Read 2 score (1–3)	2.234	0.174	2,460	0.063	3.258	0.254	2,460	0.063
Read 2 score - Read 3 score	0.514	0.168	2,460	0.004	0.750	0.245	2,460	0.004
Admission	0.839	0.113	2,460	0.022	1.223	0.165	2,460	0.022

Notes: Each row reports two bivariate regressions. The dependent variable is as indicated on the left. The independent variable in columns 1–4 is the random forest prediction of whether the second reader's score will be higher than the third reader's score, given the letter features. Columns 5–8 rescale that prediction to have coefficient 1 in the test sample for that outcome.

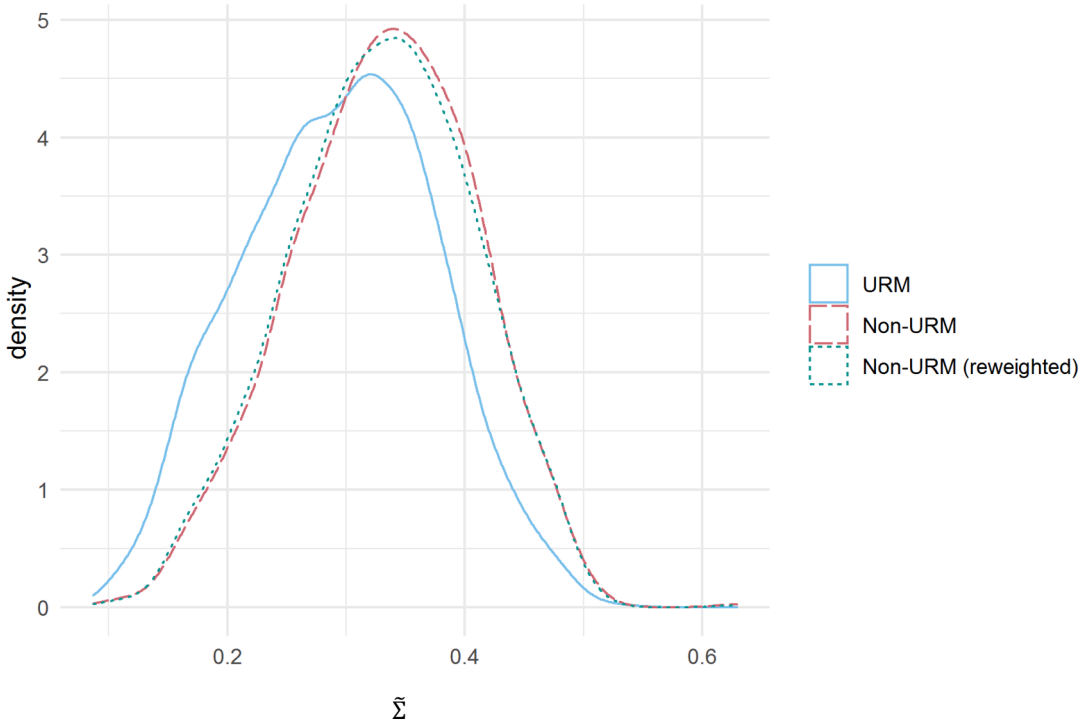

Fig. 3. Distribution of scaled letter strength ($\tilde{\Sigma}$) among URM and non-URM applicants in experimental sample
Notes: Figure shows the density of the rescaled letter strength index, the predicted probability that reader 2, with access to the letters, will give a higher score than reader 3, without access, given the letter features, and is estimated only from the subsample who do not receive "Yes" scores from reader 3. The "reweighted" series reweights non-URM students to have a similar distribution of the Admissions Index (AI) as the URM students.

Table 8
Predicting letter strength.

	(1)	(2)	(3)	(4)	(5)
URM	−0.026	−0.022	−0.015	−0.013	−0.004
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
Academic Index		0.042		0.026	0.016
		(0.005)		(0.006)	(0.006)
Read 1: Possible			0.024	0.023	0.016
			(0.003)	(0.003)	(0.003)
Read 1: Yes			0.024	0.019	0.019
			(0.006)	(0.006)	(0.006)
Read 3: Possible			0.007	0.004	0.003
			(0.003)	(0.003)	(0.003)
Read 3: Yes			0.012	0.006	0.004
			(0.004)	(0.004)	(0.004)
Application-based index of URM (D^X)					−0.031
					(0.005)
Letter-based index of URM, adjusted for application features (Δ^W)					−0.160
					(0.018)
R^2	0.053	0.079	0.100	0.108	0.143

Notes: Dependent variable is $\tilde{\Sigma}_i$, the letter strength index rescaled to an unbiased predictor for the test sample. $N = 2{,}460$.

adds controls for the reader-1 and reader-3 scores, each conducted without benefit of the letters, and column 4 includes both. The reader scores might capture aspects of the application's strength that are available even without the letters but difficult to proxy with quantifiable application characteristics. Their inclusion reduces the gap in $\tilde{\Sigma}_i$ by close to half, but it remains substantial. Column 5 adds the application distinctiveness scores from Section III. This eliminates the gap in letter strength, with most of the work being done by the letter-based measure. Evidently, the distinctiveness indices are capturing something that is valued by application readers – even though I had difficulty distinguishing letters written for URM and non-URM students, the index I obtained has meaningful signal, and readers reward letters that look like those written for non-URM students. It is impossible to tell from the available data whether this is something that is *rightly* valued, or whether the readers are unreasonably discriminating against students from URM groups. Moreover, although Δ^W is a strong predictor of $\tilde{\Sigma}$, its very limited variability means that it does not account for much variation in letter strength: A standard deviation change in Δ^W accounts for only 17% of a standard deviation of $\tilde{\Sigma}$.

6. How does inclusion of letters affect the evaluation of applications?

As a final analysis, I investigate the relationship between my constructed letter strength measure and admissions outcomes, and in particular how the inclusion of letters affects URM-non-URM differences in these outcomes. I estimate linear probability models for two outcomes: The second reader's score, treated as a numeric measure ranging from 1 to 3, and admissions outcome, treated as binary.

Column 1 of Table 9 presents models for these outcomes that include just $\tilde{\Sigma}_i$, the letter strength index. Students with stronger letters receive much stronger scores from reader 2 and are much more likely to be admitted. The standard deviation of $\tilde{\Sigma}_i$ is 0.057, so a one-standard deviation increase in $\tilde{\Sigma}_i$ is associated with an increase of 0.19 in the reader 2 score (that is, about one-fifth of the distance from a Possible to a Yes) and a 7.0 percentage point increase in the probability of admission.

Strong letters are partially reflective of already strong applications, so this much overstates the impact of the letters themselves. Column 2 adds controls for the AI and the 1st and 3rd reader scores. This reduces the coefficient on $\tilde{\Sigma}_i$ by more than two-thirds; it remains statistically significant and practically large in the models of reader scores, but is only marginally significant in the model for admissions.

Columns 3–5 present models that omit the letter index and include

Table 9
Letter strength and application outcomes.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Dependent variable is Read 2 score								
$\widetilde{\Sigma}_i$	3.258	0.905				0.946	1.521	1.485
	(0.254)	(0.207)				(0.208)	(0.333)	(0.335)
URM			−0.293	−0.159	0.036	0.049	0.066	0.103
			(0.029)	(0.027)	(0.026)	(0.026)	(0.028)	(0.041)
URM*$\widetilde{\Sigma}_i$							−0.938	−0.795
							(0.425)	(0.436)
URM*AI							0.024	0.109
							(0.112)	(0.122)
URM*D^X								0.198
								(0.111)
URM*Δ^W								0.245
								(0.393)
AI		y		y	y	y	y	y
Reader 1 & 3 scores		y			y	y	y	y
D^X, Δ^W								y
R^2	0.063	0.442	0.040	0.230	0.438	0.442	0.444	0.444
Panel B. Dependent variable is an indicator for an admission offer								
$\widetilde{\Sigma}_i$	1.223	0.245				0.326	0.499	0.532
	(0.165)	(0.152)				(0.152)	(0.244)	(0.245)
URM			−0.036	0.037	0.092	0.096	0.102	0.086
			(0.019)	(0.018)	(0.019)	(0.019)	(0.021)	(0.030)
URM*$\widetilde{\Sigma}_i$							−0.282	−0.145
							(0.312)	(0.319)
URM*AI							−0.005	0.060
							(0.082)	(0.089)
URM*D^X								0.120
								(0.081)
URM*Δ^W								0.519
								(0.287)
AI		y		y	y	y	y	y
Reader 1 & 3 scores		y			y	y	y	y
D^X, Δ^W								y
R^2	0.022	0.247	0.002	0.140	0.254	0.255	0.255	0.262

Notes: $N = 2{,}460$. All controls included in interactions with URM are measured relative to their average in the URM subsample.

URM controls. URM students receive much lower reader 2 scores than non-URM students, on average, and are 3.6 percentage points less likely to be admitted. The reader 2 gap shrinks by half when the AI is controlled, while the admissions gap is reversed. When I add controls for the 1st and 3rd reader scores – though note that the reader scores could themselves be influenced by bias for or against URM students – URM students do better on admissions and reader scores (the latter not statistically significant) than non-URMs with the same AI and 1st and 3rd reader scores. Because the 2nd reader observes the letters of recommendation but the 1st and 3rd readers do not, this indicates that URM students are relatively advantaged by the inclusion of letters in the application review.

Column 6 includes both the URM indicator and the letter strength measure. URM students do better than non-URM students with the same AIs, 1st and 3rd reader scores, and letter strength index. Column 7 adds interactions between URM and the AI and the letter strength measure. For ease of interpretation, both the AI and $\widetilde{\Sigma}_i$ are normed relative to their mean in the URM applicant pool, so the URM coefficient can be interpreted as the difference in outcomes for URM students with average values of the AI and $\widetilde{\Sigma}_i$.

The URM - $\widetilde{\Sigma}_i$ interaction is negative, indicating that readers place much less weight on a strong letter for URM than for non-URM applicants. The URM-AI interaction is near zero, as expected: Any differences in the way that readers interpret the non-letter components of the applications of URM and non-URM applicants should be absorbed by the 1st and 3rd reader scores. (When I omit these controls, the URM- AI interaction is large and positive.) The URM main effect remains positive and significant.

The results in column 7 bear some further explication. They indicate that a URM applicant with average AI and $\widetilde{\Sigma}_i$ tends to receive higher reader scores and to be more likely to be admitted than a non-URM applicant with the same AI and letter strength score, but that as letter strength improves, the URM advantage is reduced. Thus, the inclusion of letters appears to help the average URM student, but the features that are generally indicative of strong letters do not seem to convey the same advantage for URM as for non-URM applicants.

One potential explanation for this is that letters convey multiple pieces of information about a student's potential. For example, one might imagine that readers extract from the letters information both about the challenges that a student has overcome and about achievements that are not captured by the quantitative measures used to construct the AI. The former plausibly varies more between URM and non-URM applicants, while the variation in the latter may be primarily within group. It is plausible that the letter strength index $\widetilde{\Sigma}_i$ is largely measuring the achievement component of the letters and the URM main effect is capturing the "overcoming challenges" component. One way to assess this is to bring in the letter distinctiveness information from Section III. In column 8, I add controls for both D^X and Δ^W, both on their own and interacted with URM. These have basically no effect on the model for reader 2 scores, for which the R2 is unchanged to three decimal places. They have a bigger effect in the model for admission: Letters typical of URM candidates are associated with substantial admissions advantages for URM students. Controlling for this reduces the URM-by-$\widetilde{\Sigma}_i$ interaction by half: Once I allow for the fact that URM-distinctive letters are rewarded more heavily for URM applicants, I no longer find much evidence that *stronger* letters are rewarded differentially. We still see a substantial positive URM main effect – the average URM applicant receives a substantial bonus from the inclusion of letters in the evaluation.

Overall, this pattern of results seems consistent with the above explanation that letters convey information about obstacles overcome as

Appendix Table 1. Custom-built topical dictionaries.

Topic	Included words
Academic	AP, GPA, honor, academ*, learn, ability, course, class, grade, research, rigor, curriculum, tutor, merit, stud, intellect*, brilliant, analy*, library, graduat*, undergrad*, colleg*, schola*, sophisticat*, seminar, logic, bright, smart, test, professor, advanced placement, critical think, top student, curious
STEM	scien*, math, calculus, chemi*, physics, biolog*, engineer, lab, computer, tech, algebra, laborator*, stem, programm*, experiment, robot, statistic, code, geometry, trigonometry, software, physiology, coding
Humanities	writer, histor*, english, language, litera*, essay, philosoph*, poet, poem, rhetoric, latin
Arts	music, art, perform, violin, piano, cello, viola, choir, sing, paint, orchestra, chorus, theater, theater, drama, actor, actress, visual, band, flute, oboe, clarinet, bassoon, horn, trumpet, trombone, tuba, harp, saxophone, percussion, sculpture, dance, talent, film, stage, painter
Social sciences	psychol*, econom*, business, govern*, politic*, sociolog*, journalis*
Languages	spanish, french, german, chinese, japanese, korean, thai, mandarin, cantonese, russian, india, hindi, farsi, english language
Extracurriculars	club, leader, team, group, president, extracurricular, led, organiz*, association, participat*, involv*, director, join, camp, officer, newspaper, yearbook, trial, elect*, scout, secretary, treasurer, nations
Sports	coach, team, sport, athlet*, olymp*, practice, train, football, volleyball, basketball, soccer, gymnastics, lacrosse, baseball, wrestl*, swim, diver, diving, run, compet*, tournament, champion, varsity, JV, tennis, game, decathalon, golf, polo, cheerlead*, rugby, track, cross country
Family/ community	communit*, member, group, family, service, classmate, social, volunteer, coordinate, society, collaborat*, parent, brother, sister, mom, dad, grandparent, friend, peer, together, belong, fellow, children, church, religious, assist, cultur, sibling, partner, voluntar*, support, colleague, nonprofit, grandmother, grandfather, mother, father, help others, helping others
Grit	dedicat*, challeng*, commit*, difficult*, goal, motivat*, determin*, drive, enthusias*, confiden*, focus, effort, practice, train, compet*, achiev*, initiative, accomplish, effort, persever*, demanding, pressure, obstacle, work, persist, pursue, ambitio*, tenaci*, resilien*, push, prove, hard, strive, adversity, grit

Notes: All words beginning with the indicated stem were counted; asterisks indicate common stem words. In some cases (e.g., "help others"), only the indicated two-word sequence was counted.

well as applicant achievement, and that the average URM applicant benefits from the former.

7. Conclusion

There are many aspects of selective admissions that create explicit or implicit advantages for students from high-scoring high schools and wealthy families who have experience with the college application process. Efforts to increase fairness in admissions amount to increasing or reducing the weights given to particular measures, in the hope that the fairer measures can be given the most weight. But measuring the impact of any particular measure on admissions is not straightforward.

I use data from a pilot study conducted by the University of California, Berkeley, in 2017. A subset of applicants that year were asked to submit letters of recommendation from two teachers, counselors, or other adults. Working with the admissions office, I obtained the text of the letters submitted. I also oversaw a within-subjects experimental design, wherein applications of students who submitted letters were re-read under identical conditions but without the letters. I use comparisons between application scores given by readers who had the letters and those from the parallel no-letter evaluations to identify which letter features lead to higher application scores.

I show that letters written on behalf of underrepresented applicants are only minorly different from those written on behalf of other applicants with similar quantitative credentials. One exception to this general pattern is that students from low-scoring high schools do seem to receive different letters, perhaps indicative of a shortage of adults at these schools who can write strong letters.

I also show that it is possible to identify features of letters that lead to stronger application scores. Although my use of natural language processing to identify letter strength leads to a "black box" estimate, features characterizing the degree to which the letters emphasize academic strength appear important in determining reader scores.

Last, I show that underrepresented applicants achieve better average application outcomes – reader scores, admissions decisions – when letters are available. These gains are *not* concentrated among the students with the strongest letters, according to the above index; rather, students with average letters benefit more than do those with stronger letters. A plausible interpretation, consistent with the data, is that readers infer two different dimensions of applicant characteristics from the letters, both of which they reward. They infer applicants' academic strength, captured by the letter strength index, and they infer the degree to which the applicant has overcome barriers. The latter is widely seen to be an important qualification for selective colleges, but it is not easy to observe even in holistic review systems like that in place at UC Berkeley. There is a case for including subjective information like letters in the process in order to make it more visible, at least within systems like Berkeley's that are carefully designed to promote equitable admissions.

Appendix to Qualitative Information in Undergraduate Admissions: A Pilot Study of Letters of Recommendation

As discussed in the text, the analysis was conducted on deidentified data. This appendix describes the deidentification process in more detail.

The analysis combines traditional data from administrative records maintained by the admissions office, with variables like GPAs, SAT scores, demographics, and readers' assigned scores, with data derived from the text of the letters of recommendation. The administrative records are deidentified before being provided to the research team – individual identifiers like names and dates of birth are replaced with a coded ID. The process for providing matching versions of the letters involved several steps:

1 PDF files of letters were extracted to generate an unstructured text field containing the full contents of the PDF file.
2 A new variable was created containing the coded ID of the student for whom the letter was submitted. Additional variables were created with other metadata, most notably the student's description of his or her relationship with the letter writer.
3 The unstructured text field was modified to remove the first name and last name of the student and of the letter writer, wherever they appeared.
4 The text field was further cleaned to select only the text between the salutation and the closing. This was intended to remove addresses as well as artifacts from the letter writer's letterhead.

Steps 1–3 were conducted by the admissions office. The resulting file was passed to researchers, who conducted step 4.

References

Alvero, A. J., Arthurs, N., Antonio, A. L., Domingue, B. W., Gebre-Medhin, B., Giebel, S., et al. (2020). AI and holistic review: Informing human reading in college admissions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 200–206). https://doi.org/10.1145/3375627.3375871

Alvero, A. J., .Giebel, S., Gebre-Medhin, B., Antonio, A. L., .Stevens, M. L., & Domingue, B. W. (2021). Essay Content is Strongly Related to Household Income and SAT Scores: Evidence from 60,000 Undergraduate Applications. Stanford Center for Education Policy Analysis working paper 21-03.

Atkinson, R. C. (2001). The 2001 Robert H. Atwell distinguished lecture. In *83rd Annual Meeting of the American Council on Education*. University of California.

Bastedo, M. N., Bowman, N. A., Glasener, K. M., & Kelly, J. L. (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), 782–805. https://doi.org/10.1080/00221546.2018.1442633

Bendick Jr,, M., & Nunes, A. P. (2012). Developing the research basis for controlling bias in hiring. *Journal of Social Issues, 68*(2), 238–262.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991–1013.

Biernat, M., & Fuegen, K. (2001). Shifting standards and the evaluation of competence: Complexity in gender-based judgment and decision making. *Journal of Social Issues, 57*(4), 707–724.

Hoxby, C. & Turner, S. (2013).Expanding college opportunities for high-achieving, low income students. Stanford Institute for Economic Policy Research Discussion Paper. https://siepr.stanford.edu/sites/default/files/publications/12-014paper_6.pdf.

Bleemer, Z.. Top percent policies and the return to postsecondary selectivity. Working paper. http://zacharybleemer.com/wp-content/uploads/2020/10/ELC_Paper.pdf.

Bleemer, Z. (2022). Affirmative Action, Mismatch, and Economic Mobility After California's Proposition 209. *Quarterly Journal of Economics, 137*(1), 115–160.

Bowen, W. G., & Bok, D. (1996). *The shape of the river: Long-term consequences of considering race in college and university admissions*. Princeton, N.J.: Princeton University Press.

Cohodes, S. R., & Goodman, J. S. (2014). Merit aid, college quality, and college completion: Massachusetts' Adams Scholarship as an in-kind subsidy. *American Economic Journal: Applied Economics, 6*(4), 251–285. https://doi.org/10.1257/app.6.4.251

Black, S. E., Denning, J. T., & Rothstein, J. (2021). Winners and losers? The effect of gaining and losing access to selective colleges on education and labor market outcomes. Working paper, June. https://eml.berkeley.edu/~jrothst/workingpapers/top_ten_06_2021.pdf.

Ben-Michael, E., Feller, A., & Rothstein, J. (2021). Varying impacts of letters of recommendation on college admissions: Approximate balancing weights for subgroup effects in observational studies. Working paper, August. https://eml.berkeley.edu/~jrothst/workingpapers/BMFR_LOR_aug2021.pdf.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology, 92*(6), 1006–1023.

College Board (2020). *Understanding SAT scores*.https://collegereadiness.collegeboard.org/pdf/understanding-sat-scores.pdf.

Gelman, A., Fagan, J., & Kiss, A. (2007). An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association, 102*(479), 813–823.

Glaser, J., Martin, K. D., & Kahn, K. B. (2015). Possibility of death sentence has divergent effect on verdicts for Black and White defendants. *Law and Human Behavior, 39*(6), 539–546.

Houser, C., & Lemmons, K. (2018). Implicit bias in letters of recommendation for an undergraduate research internship. *Journal of Further and Higher Education, 42*(5), 585–595. https://doi.org/10.1080/0309877X.2017.1301410

Hout, M. (2005). *Berkeley's comprehensive review method for making freshman admissions decisions: An assessment*. Berkeley Academic Senate. https://academic-senate.berkeley.edu/sites/default/files/hout_report_2005.pdf.

Hoxby, C., & Avery, C. (2013). The missing "one-offs": The hidden supply of high-achieving, low-income students. *Brookings Papers on Economic Activity, 2013*(1), 1–65. https://doi.org/10.1353/eca.2013.0000

Cortes, K., & Klasik, D. (2020). Uniform admissions, unequal access: Did the top 10% plan increase access to selective flagship institutions? National Bureau of Economic Research working paper no. 28280. http://www.nber.org/papers/w28280.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer. Springer Texts in Statistics.

Karabel, J. (2005). *The chosen: The hidden history of admission and exclusion at harvard, yale, and princeton*. Boston, M.A.: Houghton Mifflin Harcourt.

Krueger, A., Rothstein, J., & Turner, S. (2006). Race, income, and college in 25 years: Evaluating Justice O'Connor's conjecture. *American Law and Economics Review, 8*(2), 282–311. https://doi.org/10.1093/aler/ahl004

Kurlaender, M., Reber, S., & Rothstein, J. (2020). *UC regents should consider all evidence and options in decision on admissions policy*. Policy Analysis for California Education. https://edpolicyinca.org/newsroom/uc-regents-should-consider-all-evidence-and-options-decision-admissions-policy.

Lai, Y. L., & Zhao, J. S. (2010). The impact of race/ethnicity, neighborhood context, and police/citizen interaction on residents' attitudes toward the police. *Journal of Criminal Justice, 38*(4), 685–692.

Lemann, N. (1995). The Great Sorting. *The Atlantic Monthly, 276*(3), 84–100.

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology, 94*(6), 1591–1599.

Oliveira, A., & Murphy, K. (2015). Race, social identity, and perceptions of police bias. *Race and Justice, 5*(3), 259–277. https://doi.org/10.1177/2153368714562801

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2009). Does unconscious racial bias affect trial judges? *Notre Dame Law Review, 84*(3), 1195–1246.

Rothstein, J. (2004). College performance predictions and the SAT. *Journal of Econometrics, 121*(1–2), 297–317. https://doi.org/10.1016/j.jeconom.2003.10.003

Rothstein, J. (2017). *The impact of letters of recommendation on UC Berkeley admissions in the 2016-17 cycle*. California Policy Lab. https://www.capolicylab.org/wp-content/uploads/2017/07/Recommendation-Letters-UC-Berkeley-July-11-2017.pdf.

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex roles, 57*(7–8), 509–514. https://doi.org/10.1007/s11199-007-9291-4

Stevens, M. L. (2007). *Creating a class: College admissions and the education of elites*. Cambridge, MA: Harvard University Press.

US News (2021). *University of California Berkeley*.https://www.usnews.com/best-colleges/university-of-california-berkeley-1312.

Weinberg, J. D., & Nielsen, L. B. (2012). Examining empathy: Discrimination, experience, and judicial decisionmaking. *Southern California Law Review, 85*(2), 313–351.