NBER WORKING PAPER SERIES

SUBJECTIVE EVALUATIONS AND STRATIFICATION IN GRADUATE EDUCATION

Jessica Bai
Matthew Esche
W. Bentley MacLeod
Yifan Shi

Working Paper 30677
http://www.nber.org/papers/w30677

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2022

We have no outside funding for this project. The authors thank Janet Currie, Racquel Fernandez, Matt Jackson and Miguel Urquiola for helpful discussions and comments. We also thank participants at the Berkeley labor seminar for helpful comments. The support of the Program for Economic Research is gratefully acknowledged. Corresponding author: W. B. MacLeod, wbmacleod@ wbmacleod.net. The project was reviewed and approved under Columbia IRB protocol AAAS2371. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Subjective Evaluations and Stratification in Graduate Education
Jessica Bai, Matthew Esche, W. Bentley MacLeod, and Yifan Shi
NBER Working Paper No. 30677
November 2022
JEL No. I24,J01

ABSTRACT

We introduce a model of the admissions process based upon standard agency theory and explore its implications with economics PhD admissions data from 2013-2019. We show that a subjective score that aggregates subjective ratings and recommendation letter features plays a more important role in determining admissions than an objective score based upon graduate record exam (GRE) scores. Subjective evaluations by references who write multiple letters are not only more influential than those of references who write one letter, but they are also more informative. Since multiple-letter references are also more highly ranked economists, this implies that there is a constraint on the supply of high-quality references. Moreover, we find that both the subjective and objective scores are correlated with job placement at a top economics department after the completion of the PhD. These indicators of individual achievement have a smaller effect than an undergraduate degree from an Ivy Plus school (i.e., Ivy League + Stanford, MIT, Duke, and Chicago). In the self-selected pool of applicants, Ivy Plus graduates are twice as likely to be admitted to a top 10 graduate program and are much more likely to obtain an assistant professor position at a top 10 program upon PhD completion. Given that Ivy Plus students must pass a stringent selection process to gain admission to their undergraduate program, we cannot reject the hypothesis that admission committees use information efficiently and fairly. However, this also implies that there may be a return to attending a selective undergraduate program in order to be pooled with highly skilled individuals.

Jessica Bai
Department of Economics
Harvard University
Cambridge, MA 02138
jbai@g.harvard.edu

Matthew Esche
Department of Economics
Columbia University
420 West 118th Street, MC 3308
New York, NY 10027
matthew.esche@columbia.edu

W. Bentley MacLeod
Department of Economics
Columbia University
420 West 118th Street, MC 3308
New York, NY 10027
and NBER
wbmacleod@wbmacleod.net

Yifan Shi
Department of Economics
Columbia University
420 West 118th Street, MC 3308
New York, NY 10027
ys3094@columbia.edu

1. Introduction

The canonical labor market model assumes that employment and compensation are set by competitive market clearing (Acemoglu and Autor (2011)). Yet, in practice, the matching of individuals to jobs is a very complex process, particularly in the market for highly skilled young workers. Recruiting costs make it impossible for an employer to evaluate all prospective hires. Rather, matching begins with prospective hires self-selecting to apply to a limited number of potential employers who then evaluate the prospective hires based upon the information provided.

This evaluation process typically uses both objective and subjective information. By objective information, we mean verifiable performance measures such as test scores or examples of output produced by the applicant. By subjective information, we mean assessments provided by individuals who assess the future potential of an applicant. Given this information, prospective employers then predict which of the applicants will perform the best in the future. This is necessarily a noisy and imperfect process that can lead to compressed starting wages and high turnover early in a person's career.[1] Moreover, the use of subjective evaluations implies that applicants' characteristics that are unrelated to performance, such as their gender or nationality, may affect their success on the job market.

One goal of this paper is to better understand the role of subjectivity in the evaluation process by exploiting the highly structured environment of admissions to a PhD program in economics. A PhD is a necessary requirement for most academic positions in economics. It may also lead to employment in both the public and private sectors outside of academia. The fact that there are normally five to six years between admission to a graduate program and initial employment implies a great deal of uncertainty regarding future success at the time of graduate school. The question then is, what factors best predict future success? In this paper, we introduce a framework for studying this process. We provide evidence on how applicants to a top graduate program in economics are evaluated, and subsequently, how these evaluations are related to academic job placements.

Our data consists of 6,320 applications to a top PhD program in economics between 2013 and 2019.[2] Using hand-collected data on applicant outcomes, we match each of the program's applicants to publicly observed admissions outcomes. We also match applicants from 2013-2015 to assistant professor job outcomes. Approximately 20% of the sample attends a top ten program and 4% of the 2013-2015 sub-sample attains a toptwenty assistant professor position. Each year, the sample includes between 800 to 1,000 applications with a rich set of demographic information, test scores, and evaluations from at least three references.

We begin by adapting the insights from agency theory to organize our results and illustrate how incentives interact with selection.[3] Given the limited number of slots available in a PhD program, admissions committees must develop criteria to select individuals from the pool of applicants. This can be viewed as a principal-agent problem in which the admissions committee (the principal) offers a reward to person i (the agent), $a_i = 1$ if admitted and $a_i = 0$ if not, as a function of their observed

[1] A classic paper in this literature is Topel and Ward (1992) which documents early career patterns in the US.
[2] The graduate school supplying this data has asked us not to reveal its identity.
[3] See Holmström (1979) and Harris and Raviv (1979) who emphasize the importance of information for agency theory.

performance in their application. A natural hypothesis is that programs wish to admit individuals who are most likely to succeed at a career in economics. We can let the probability of success of individual i be given by p_i, and suppose that it is related to some measure of skill, $\hat{\alpha}_i$, formally defined by $p_i = p(\hat{\alpha}_i)$. This does not mean that skill is one-dimensional, but that the various characteristics of a person can be combined into a single index that can be used to rank applicants as a function of their expected future success.

We use admission to any top ten economics program as the outcome or reward. Here we assume that admissions committees have similar preferences and share a desire to admit students who are likely to have successful academic careers. A documented feature of the economics profession is the existence of shared values regarding quality (see Fourcade et al. (2015)). This consensus assumption is that all the programs rank applicants similarly, with differences arising from idiosyncratic variation in their assessments. One of our findings is that the revealed preferences of admissions committees over applicants are consistent with the applicants' success on the job market when they complete their program.

The hypothesis that admissions committees' rankings and the final ranking of applicants on the academic job market are consistent with each other has several empirical implications that we detail in section 3. First, potential applicants have access to their graduate record exam (GRE) results before they apply. Given that admission is not certain, the theory predicts that potential applicants trade off the chances of admission against the cost of applying, and the benefits that would accrue if accepted. This implies that there should be more applications from individuals who believe that they have a greater chance of admission. If applicants expect GRE scores to be correlated to admissions chances, then this should be observed in the data. The distribution of quantitative GRE scores is illustrated in figure 1.

Since we plot the density by percentile score, if the sample included a random sample of all test takers, then the distribution would be uniform. Given this is not the case, there is clear evidence that individuals self-select as a function of their quantitative GRE scores. Interestingly, the distribution of the verbal and writing scores, illustrated in the appendix, appears to be less highly selected, and consistent with the fact that the applicant pool consists of many non-native English speakers.

This self-selection implies that one cannot necessarily reject the hypothesis of random admissions. If slots were randomly allocated, then the GRE distribution of admitted students would be given by the black line, and the sample would *appear* selected. Since we do not have experimental variation, it is impossible to identify the admissions policy for an arbitrary distribution of applicants. It is worth highlighting that since the space of possible applicant distributions is infinite-dimensional, a purely experimental approach to measuring committee behavior would be impossible. Some theory is needed to constrain the set of potential hypotheses.

There are two empirical questions we can answer. The first is that we can reject the hypothesis of random admissions. Figure 1 also plots the density of admitted students in blue, where one can see direct evidence of positive selection by GRE. Individuals with higher GRE scores are more likely

to be admitted.[4] It is sometimes claimed that committees use cutoff rules, notice that while there is evidence of selection as a function of the GRE score, there does not appear to be a strict cutoff rule. We observe some applicants with scores less than the 70th percentile being admitted to some top ten programs.

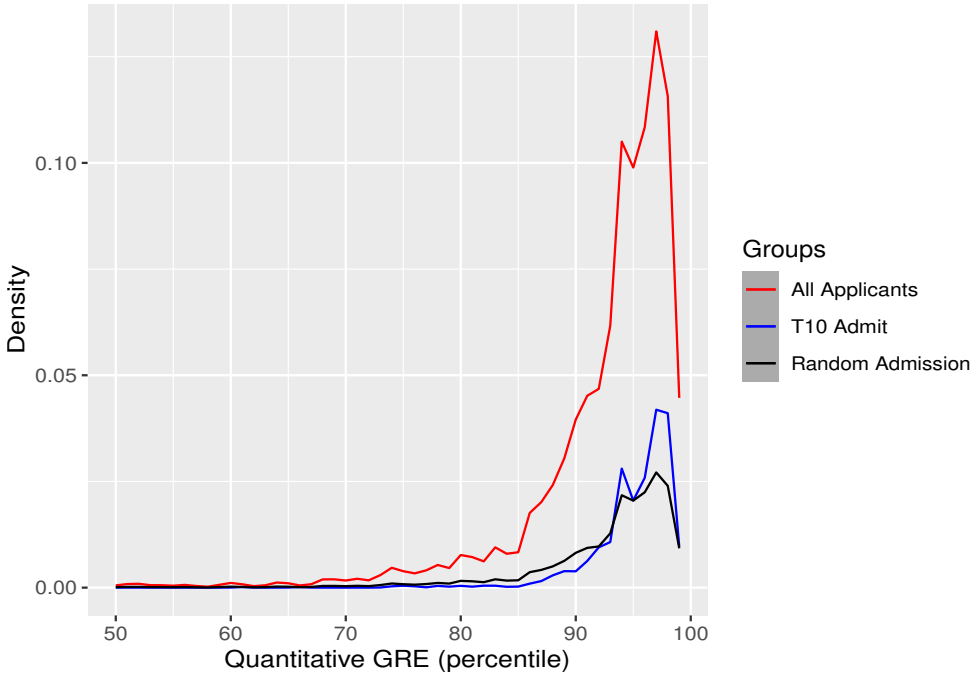


FIGURE 1. Distribution of GRE Quantitative GRE Scores in the Sample

The second question uses the fact that admissions are not random to explore factors that determine admissions other than GRE scores. Each applicant submits at least three references. A reference submits a letter of recommendation and provides four subjective ratings on a five-point scale (see figure 3). This information is not observed by the applicant, and hence the applicant cannot use it directly to make their application decision. Building upon previous work, we scan the letters of recommendation to identify the sentiment in a letter that explains potential research performance (see discussion of the literature in the next section). The ratings and letter sentiments data form the basis for a *subjective score*. The three GRE scores from the quantitative, verbal, and writing exams form the basis for the *objective score*. We assess the relative importance of these measures for admissions and subsequent job placement as well as how they are related to the gender and citizenship of the applicants.

We find, consistent with the predictions of agency theory, that all sources of useful information are used to assess applicant quality. However, the subjective score based upon reference information is significantly more important (in terms of standard deviation effects) than the objective score based upon GRE ratings. Moreover, while both objective and subjective scores are correlated with

[4]Note that there is a slight dip at the top of the distribution. This is generated by top scoring on exams and how the Educational Testing Service (ETS) assigns percentiles.

individuals obtaining a top ten assistant professor position, after controlling for the ranking of the PhD program a person attends, only the subjective score remains significant.

These results are consistent with Athey et al. (2007) who find little effect of the GRE score on job placement. We also find that the subjective evaluations by the references at the time of admission are positively correlated with job market success five to six years after admission to a top-ten graduate program. This is true even though the subjective score is very noisy. We can see this by comparing the ratings of the same applicant by different references. (see figures 10-11). Agency theory also predicts that one can improve upon information quality by using relative performance measures (Green and Stokey (1983)). Consistent with this prediction, we find that the relative subjective score is positively correlated with admissions.

The Spence (1973) signaling model predicts that effort can be used as a screening device. In this case, we would expect that a reference can signal applicant quality by writing a longer letter. We find no evidence that *relative* letter length is correlated with admissions chances. However, we do find that letter length is associated with a higher admissions rate. This would be consistent with admissions committees using length as a signal, even though there is little evidence that individual references vary letter length to signal the quality of their applicants.

We follow Stock and Siegfried (2014), and document the relationship between admissions and identity, as measured by citizenship and gender.[5] Goldin and Rouse (2000) find that removing gender barriers may decrease the quality of the pool of female applicants, who subsequently believe that their employment chances have improved. We find that the pool of female applicants has, on average, lower quantitative GRE scores than their male counterparts, consistent with this hypothesis. However, after controlling for observed quality, we do not find that identity has a significant impact on admission success.

Aside from measures of personal performance, we find that there are two signals in an individual's application that have a large and significant relationship with an applicant's admissions success. The first of these is the number of letters a reference writes for people in our applicant pool. Individuals who have references that are prolific letter writers in our sample have a significantly greater chance of admission. This is a fixed effect based on the *identity* of the reference.

Not surprisingly, prolific letter writers tend to be more productive researchers. In addition, their subjective evaluations are more discriminating. There is no payment for providing a reference; moreover, there is no feedback regarding how references rate applicants. Hence, references are free to top-code applicants by giving them the highest possible subjective grade on the subjective scale references are asked to fill. However, we find that references who write multiple letters tend not to do this. In particular, there is a steeper relationship between applicants' objective scores and the subjective scores from references who write several letters than ones who write a single letter. This implies that multiple letter writers are more discriminating, and hence a rational decision-maker using Bayesian updating would put more weight upon their evaluation.

[5]Both of these characteristics are measured for all applicants. We have data on race for US applicants only. That field is, however, voluntary. In the end, less than 0.8% of our sample is a part of an underrepresented minority group. Thus, we do not have the power to explore the interaction between race and applicant evaluation.

The limited supply of such individuals implies a capacity constraint on obtaining influential letters. This capacity constraint is also evident with another signal available to admissions committees, namely whether or not the applicant attended an elite undergraduate program. We find that attending an Ivy Plus undergraduate school has a signifcant positive effect on admission chances, consistent with college quality signaling skill (Spence (1973); MacLeod and Urquiola (2015)).[6]

The agency approach provides a framework to understand these phenomena. The fact that the number of applicants greatly exceeds a program's capacity constraints makes it very expensive to evaluate individual performance. The fact that evaluation is expensive has motivated decades of research into designing better tools to measure skill (see Ployhart et al. (2017)). If the evaluations of the undergraduate admissions committees are highly correlated with the goals of graduate admissions, there may be statistical discrimination—the graduates of Ivy Plus colleges are more highly selected, and hence as MacLeod and Urquiola (2015) predict, they should do better in the market for graduate admissions. This arises from the combination of costly evaluation and the existence of common evaluation criteria for acceptance to an Ivy Plus program.

If the Ivy Plus signal is non-informative, then one might expect that the disproportionate admission of Ivy Plus applicants to top graduate programs would lead to proportionally lower success rates in the job market. In fact, consistent with the evidence in Stock and Siegfried (2014), we find Ivy Plus applicants are much more successful in obtaining assistant professor positions at top-ten programs. Hence, from the perspective of the admissions committee, the overweighting of attending an Ivy Plus school is consistent with the hypothesis that they wish to admit individuals who are successful in the academic job market.

Thus, our analysis is consistent with a structure in which the goals of both the admissions committees and the hiring committees on the academic market are correlated. With this data, we are not in the position to conduct a normative analysis of this process. Rather, as we discuss further in the concluding discussion, we highlight how agency theory can provide a useful approach to understanding admissions data by highlighting the revealed preferences of the parties in the process, and how evaluating parties weigh the various sources of information regarding the performance of applicants.

The agenda for the paper is as follows. The next section discusses the relevant literature, followed by a discussion of the agency framework used to organize the results. Section 4 discusses our data, while section 5 outlines the machine learning procedure used to produce the objective and subjective scores. Section 6 documents the relative importance of these scores for admissions and the extent to which they are correlated. Section 7 documents the within-applicant and within-reference reliability of the subjective scores. Section 8 presents the results on job placement for the subset of the individuals where job market outcome data is available. Our concluding discussion in section 9 discusses the implications of our results for hierarchical stratification.

[6]Ivy Plus includes the Ivy League schools plus Stanford, Duke, University of Chicago, and MIT.

2. Literature

This paper contributes to both the literature on admissions in educational settings and on performance evaluations in labor markets. Within the admissions literature, we contribute to the small literature focusing on economics graduate programs and to a broader literature that uses statistical models to aid and describe admissions processes. We also join a literature focused on documenting performance evaluations, including subjective evaluations in the workplace and the use of referrals and reference letters during hiring. We connect both literatures by focusing on the role of subjective evaluations in a setting where we can compare their relative importance with other evaluations, like test scores.

2.1. **Literature on Admissions.**

This paper adds to the literature that examines predictors of admission and performance using admissions data for economics graduate programs. Krueger and Wu 2000 use one year of data of an applicant pool to a top-five economics department and find that GRE scores and ratings by the admissions committee predict applicant job placement. Athey et al. 2007 assemble a dataset of 1,029 economics graduate students enrolled at top-five programs in the 1990s. They find that students who attended elite undergraduate universities and liberal arts colleges are more likely to be placed in top-ranked academic jobs. Grove and Wu 2007 find that quantitative GRE scores and reference writer prominence are positively associated with long-run publications and that admissions committee ratings predict doctoral completion and publishing.

In contrast to these studies, we examine seven years of applicant data and analyze detailed evaluations including the textual content of reference letters. We find that references' subjective ratings, a performance assessment not considered by previous studies, as well as the content of reference letters, strongly predict graduate school admission.

There is also a related literature exploring the effect of letters of reference upon admissions. Trix and Psenka (2003), examining 312 letters of recommendation for medical faculty at an American medical school, find systematic differences in recommendation letters for female and male students. Specifically, they find that letters for women tend to be shorter on average than letters for men and that a greater percentage of letters written for female applicants are "letters of minimal assurance," meaning that the letters lacked a stated commitment to the applicant. Madera et al. (2009) study 624 recommendation letters for 194 applicants for junior faculty positions from 1998 to 2006 at a southern university in the US. They find that applicant gender predicts the use of adjectives relating to the applicant's communal (e.g., terms such as affectionate, helpful, kind, and sensitive) or agentic (e.g., terms such as assertive, confident, independent, and intellectual) nature, with women more often described as communal and men more often described as agentic. Issacc et al. (2011) study 297 medical student performance evaluations for linguistic differences according to student and author gender. They find small but significant differences by gender, though these differences did not appear to lead to differences in outcomes. Dutt et al. (2016) study 1,224 letters of recommendation for postdoctoral fellows in geosciences over the period 2007 to 2012. They find that female applicants are half as likely to receive excellent versus good letters and that the gender of the letter writer does not explain this result.

2.2. **Literature on Subjective Evaluation.** This paper also contributes to a growing empirical literature that documents the use of subjective evaluations in the workplace and their relation to performance outcomes. Frederiksen et al. 2017 document that subjective performance measures are positively correlated with career outcomes such as base salaries, bonus pay, and promotions. In contrast to studies with subjective evaluations measuring on-the-job performance, we study these evaluations in the context of hiring PhD students. Selecting applicants for PhD positions can be viewed as hiring, given that PhD students are paid wages and expected to contribute to research and teaching at their academic institutions in addition to completing coursework. A unique feature of our setting is the availability of both subjective and objective measures of applicant ability. This contrasts with many studies in the personnel literature that rely on ordinal subjective rankings only. A related education literature on evaluating teacher effectiveness documents subjective evaluations by principals and their relation to objective measures such as student achievement gains (e.g., Jacob and Lefgren 2008, Murnane 1975). We find that the variation in the admissions rate is more highly correlated with the variation in subjective evaluation supplied by an applicant's references than with the objective evaluation based upon an applicant's graduate record exam (GRE) results.

This paper also relates to the literature on the value of referrals in the labor market for both hiring and performance outcomes. Recent empirical work explores the mechanisms behind the usefulness of referrals. Pallais and Sands 2016 conduct online field experiments to find that referred workers exhibit higher performance and lower turnover than non-referred workers, suggesting that referrals do contain valuable information about worker quality. Beaman et al. 2018 provide experimental evidence that high-ability workers refer higher-performing workers and become less likely to refer relatives when they are incentivized by referral performance pay. Their findings provide evidence that referrers have private information about their contacts, which may be a valuable quality of referrals to firms. A growing body of work explores wage and tenure dynamics for referred workers. Dustmann et al. 2016 use matched employer-employee data and find that referred workers earn higher wages and are less likely to leave their firms, though this effect declines with tenure in the firm. Similarly, Brown et al. 2016 find that referred workers are more likely to be hired and experience an initial wage advantage.

More recent empirical work has emerged to study the heterogeneous effects of referrals on labor market outcomes. Lester et al. (2021) find that referred workers experience differences in starting wage and tenure depending on whether they were referred by family and friends or by business contacts. Beaman et al. (2018) study a potential cost of using informal networks to overcome labor market frictions: groups that are distant from the employed may become disadvantaged. They conduct a field experiment in Malawi and find that men systematically refer fewer women even when they are capable of referring high-quality women.

Meanwhile, the literature on referrals in the specific form of reference letters is smaller but growing. Some recent work aims to uncover causal estimates of the importance of reference letters for hiring outcomes. Abel et al. 2020 find that providing letters of reference increases the probability of a response from an employer and that letters are also valuable to employers in helping them select applicants of higher ability. In contrast to their study, recommendation letters in our setting

are required for the application. Our paper documents that the private information provided by subjective evaluations plays a key role in explaining observed features of graduate admissions, illustrating a need to better understand the relationship between subjective evaluations and graduate school performance.

Finally, there are papers in the psychology literature such as Dawes and Corrigan (1974) that make the point that statistical models of admissions can be used to both describe the admissions process and make admissions decisions.[7] There is also a growing literature in economics on the use of formal models to replace the subjective evaluations by human decision-makers (Kleinberg et al. (2018)). This literature extends the large literature in psychology that finds that formal evaluation can be superior to subjective human evaluation (Dawes (1971), Kahneman (2003) and Kahneman and Klein (2009)). Weizenbaum (1976) recommends caution in using such approaches to replace human judgment.

3. Admissions Policy: Theory

Admissions committees must allocate a fixed supply of slots to students from the pool of applicants. One method, the one required by law for charter schools, is to randomly allocate slots. However, individuals vary in their ability to pursue economics research, and hence top departments choose applicants that they believe will be successful in the profession based upon the information contained in their applications. In other words, admissions committees are engaged in a prediction problem—given the applications, which students should be given one of the limited number of slots in the program?

The answer to this question is equivalent to reward design in tournament models in agency theory, where applicants play the role of agents and admission committees play the role of principal. There are n applicants, $i \in P^0$, who should be rewarded with admission, $\delta_i = 1$, or be rejected, $\delta_i = 0$, based upon the information from the person, $\omega_i \in \Re^k, i \in P^0$. The maintained hypothesis of agency theory (Holmström (1979), Harris and Raviv (1979)) is that the principal aggregates information efficiently and then selects the most able agent to reward. In this section, we work out the implications of this hypothesis for our setting and ask what, if any, are the empirical implications of the theory.

In addition to the assumption of efficient information acquisition, it is also assumed that there is a one-dimensional measure of a person's research potential. Here we are not assuming that all individuals are the same, or have the same interests. Rather, from the perspective of the admissions committee, the main criteria is the likelihood that the person is able to get a good academic job upon graduation. Given that such a potential is not observable, we suppose that within the pool of applicants it is represented by an unobserved latent variable $\alpha_i \sim N(0,1)$. To keep matters simple, let $\hat{\alpha}_i = E\{\alpha_i|\omega_i\}$ be the conditional expectation of potential skill, and $p_i = p(\hat{\alpha}_i) \in [0,1]$ be the probability that a person will be "successful" given their latent skill $\hat{\alpha}_i$. It is assumed that $p'(\hat{\alpha}) > 0$ for $\hat{\alpha} \in \Re$. This, combined with the information on each applicant, can be used to provide predictions on how the information in an application is transformed into an offer of admission.

[7] There is also a vast applied psychology literature that we do not have room to discuss here. See Ployhart et al. (2017) for a review of a hundred years of research on the topic.

3.1. **Optimal Policy.** Suppose that the goal of admissions is to maximize the expected success of admitted applicants, given that only a fraction $r_a = m/n$ can be admitted. Let us further suppose that the information about an individual is parameterized so that $\omega_i \in \Re^k$ has density $f(\omega) > 0$ for all $\omega \in \Re^k$ for the population of individuals in the applicant pool. This can be generalized to allow for discrete distributions, though the analysis is more convenient in this case.[8] It is further assumed that $E(\alpha|\omega_i)$ is a continuous function of ω_i, $p(\hat{\alpha})$ is continuous and increasing in $\hat{\alpha}$, and $f(\omega)$ is continuous. Let the decision rule be given by $\delta \in \Delta = \left\{\delta|\delta : \Re^k \to [0,1]\right\}$, that has the interpretation that a person with application information ω_i is admitted with probability $\delta(\omega_i)$. Under this rule the payoff function is given by:

$$W(\delta) = \int_{\omega \in \Re^n} \delta(\omega)\, p(\hat{\alpha}(\omega))\, f(\omega) d\omega.$$

Thus, the admissions problem can be given by:

$$max_{\delta \in \Delta} W(\delta), \tag{1}$$

subject to the capacity constraint:

$$A(\delta) = \int_{\omega \in \Re^k} \delta(\omega)\, f(\omega) d\omega \leq r_a, \tag{2}$$

where $A(\delta)$ is the fraction of the pool that is admitted.

Proposition 1. *There exists an optimal admissions policy $\delta^*(\omega)$ and associated Lagrange multiplier, $\lambda^* \in [0,1]$, for (2) such that:*

$$A(\delta^*) = r_a,$$
$$\delta^*(\omega) \in \delta(\omega|\lambda^*), \forall \omega \in \Re^k$$

where $\delta(\omega|\lambda)$ is a correspondence defined by:

$$\delta(\omega|\lambda) = \begin{cases} 0, & y(\hat{\alpha}(\omega)) - \lambda < 0, \\ [0,1], & y(\hat{\alpha}(\omega)) - \lambda = 0, \\ 1, & y(\hat{\alpha}(\omega)) - \lambda > 0. \end{cases} \tag{3}$$

If the set of students satisfying $p(\hat{\alpha}(\omega)) = \lambda^*$ is of measure zero, then the optimal admissions rule is unique almost everywhere.

The proof of this proposition is in the appendix. This optimal decision rule has two features worth highlighting. The first is that if one's goal is to choose applicants that are likely to be the most successful in the future, then a cutoff rule is optimal. Second, the optimal rule is not necessarily unique. In particular, if at the margin there are a group of students with similar success probabilities, then any selection from this set is optimal. This point is worth highlighting because,

[8] The only difference with discrete distributions is that random admissions are optimal in some cases. The application data is sufficiently complex that randomization is a low-probability event except in the case in which signals have no information, a case we discuss in more detail below.

in the case of indifference, the program might choose other criteria for admissions, such as personal characteristics of the individuals that are unrelated to a person's future success.

3.2. **Expected Skill.** Given the goal of choosing the most promising applicants, the next step is to use the admissions data to assess applicant quality. The challenging question is how to aggregate complex information into a single performance measure that determines an applicant's rank in the pool. In particular, we wish to understand how to aggregate objective data with subjective data. By objective data, we mean the scores from the graduate record exam (GRE). These are standardized worldwide and hence provide a stable benchmark. However, there are many aspects of performance that cannot be captured in a standardized test. Thus, in addition to test scores, applicants ask at least three references to submit additional information. In order to assess the relative importance of these sources of information we assume for the moment they are represented by two signals. We combine the three GRE scores into a single index, denoted by $g_i, i \in P^0$. As is standard in the education literature, the index is z-scored within our population to have a mean of zero and a standard deviation of one. Similarly, we aggregate information from the recommendation letters and subjective ratings provided by the references to construct a subjective score, $s_i, i \in P^0$, that is also z-scored. The details of the aggregation procedure are discussed below. In this section, we discuss how committees that are interested in choosing the best applicants would use these measures.

We suppose that committees have views upon how relevant this information is regarding a person's research potential represented by the following relationships:

$$g_i = \gamma_g \alpha_i + \epsilon_i^g, \tag{4}$$

$$s_i = \gamma_s \alpha_i + \epsilon_i^s, \tag{5}$$

where $\gamma^g, \gamma^s \in [0,1]$ are measures of the admissions committee's view of the relationship between latent skill and the observed performance measure. The weights are constrained to be in $[0,1]$ to ensure the standard deviation is 1. The fact that the scores are normalized so that $var(g_i) = var(s_i) = 1$ fixes the variance of the error terms:

$$var(\epsilon_i^g) = var(g_i) - \gamma_g^2 var(\alpha_i) = 1 - \gamma_g^2,$$
$$var(\epsilon_i^s) = var(s_i) - \gamma_s^2 var(\alpha_i) = 1 - \gamma_s^2.$$

Given these signals, committees can then form beliefs regarding the latent skill of each person using Bayes' rule, summarized in the following proposition.

Proposition 2. *If (4) represents the beliefs that the admissions committees hold regarding the relationship between evaluations and potential skill, then for person $i \in P^0$, the expected skill level is given by Bayes' rule (DeGroot (1972)):*

$$\hat{\alpha}_i = E\{\alpha_i | g_i, s_i\} = \theta_g g_i + \theta_s s_i, \tag{6}$$

where the weights θ_t for $t \in \{g, s\}$ satisfy:

$$\theta_g(\vec{\gamma}) = \frac{\gamma_g\left(1-\gamma_s^2\right)}{\left(1-\gamma_g^2\right)\left(1-\gamma_s^2\right)+\gamma_g^2\left(1-\gamma_s^2\right)+\gamma_s^2\left(1-\gamma_g^2\right)} \tag{7}$$

$$\theta_s(\vec{\gamma}) = \frac{\gamma_s\left(1-\gamma_g^2\right)}{\left(1-\gamma_g^2\right)\left(1-\gamma_s^2\right)+\gamma_s^2\left(1-\gamma_g^2\right)+\gamma_g^2\left(1-\gamma_s^2\right)}, \tag{8}$$

and $\vec{\gamma} = (\gamma_g, \gamma_s)$.

Proof. Given $\alpha_i \sim N(0,1)$, and $m_{it} = \frac{g_i}{\gamma_t} = \alpha_i + \epsilon_i^t/\gamma_t$, $t \in \{g, s\}$. Then Bayes' rule implies:

$$E\{\alpha_i | m_{ig}, m_{is}\} = \frac{\rho_g m_{ig} + \rho_s m_{is}}{1+\rho_g+\rho_s},$$

where ρ_g, ρ_s are the precisions of the error terms of the signals m_{ig}, m_{is} (the inverse of the variance—$\rho = \frac{1}{\sigma^2}$). Substituting in the precisions we get:

$$\rho_t = \frac{\gamma_t^2}{1-\gamma_t^2}, t \in \{g, s\}.$$

From this we get:

$$\begin{aligned}\theta_g &= \frac{\gamma_g/\left(1-\gamma_g^2\right)}{1+\gamma_g^2/\left(1-\gamma_g^2\right)+\gamma_s^2/\left(1-\gamma_s^2\right)} \\ &= \frac{\gamma_g\left(1-\gamma_s^2\right)}{\left(1-\gamma_g^2\right)\left(1-\gamma_s^2\right)+\gamma_g^2\left(1-\gamma_s^2\right)+\gamma_s^2\left(1-\gamma_g^2\right)}.\end{aligned}$$

The expression for $\theta_g(\vec{\gamma})$ is similar. We can put our model into this form by letting $m_{ig} = g_i/\gamma_g$ and $m_{is} = s_i/\gamma_s$. Using the formulas for precision derived above and substituting into the optimal Bayes' rule yields the result. □

This result shows that the expected skill of an applicant is a linear function of the available scores. The weight placed upon these scores depends upon the beliefs $\vec{\gamma}$ regarding the information content. This proposition has a number of testable implications. The previous section shows that the optimal admissions policy is to use a cutoff rule that depends upon the number of slots available. In particular, suppose that the committees use the rule of admitting any individual with expected skill greater than $\hat{\alpha}^*$. In that case, the probability of admissions of a student with ability α_i is given by:

$$\begin{aligned}Prob[a_i = 1 | g_i, s_i] &= Prob[E\{\alpha_i | g_i, s_i\} \geq \hat{\alpha}^*] \\ &= Prob[\theta_g g_i + \theta_s s_i \geq \hat{\alpha}^*] \\ &= F\left(\frac{\theta_g g_i + \theta_s s_i - \hat{\alpha}^*}{\bar{\sigma}}\right),\end{aligned}$$

where $\bar{\sigma}^2$ is the variance of the error term. This is a standard dichotomous choice model and hence it can be easily estimated. In particular, the null hypothesis in this case would be random admissions in which one ignores the performance signals:

Corollary 3. *Suppose that it is believed that signal $t \in \{g, s\}$ has no information ($\gamma_t = 0$), then no weight is placed upon that signal ($\theta_t = 0$).*

Thus, a basic question is to simply ask if the signal provides useful information, which is simply a test of statistical significance. Of course, statistical significance does not prove that the signals are useful. If it is believed that there is no information in the signals, then there might simply be a randomization device to allocate slots. One way to think about this is to consider the relative importance of the signal. An implication of proposition 2 is that it also implies that, in the absence of perfect information, all performance signals should be used:

Corollary 4. *Suppose signal g_i has some information content but is not perfect ($\gamma_g \in (0, 1)$), then one should ignore signal g_i ($\theta_g = 0$) if and only if signal s_i has perfect information ($\gamma_s = 1$).*

This result is a version of the informative principle from Holmström (1979). The next issue is how to use this information to decide which applicants to admit. The performance measures are z-scored, which allows us to assess the relative importance of the signals via the ratio.

If signals are used optimally for admissions, then the importance of the subjective signal relative to the objective signal satisfies:

$$\frac{\theta_s}{\theta_g} = \frac{\gamma_s}{(1-\gamma_s^2)} \times \frac{\left(1-\gamma_g^2\right)}{\gamma_g}.$$

Thus, the marginal importance of the subjective signal is:

$$\frac{\partial \frac{\theta_s}{\theta_g}}{\partial \gamma_s} = \frac{(1/\gamma_s + \gamma_s)}{(1-\gamma_s^2)} \times \frac{\theta_s}{\theta_g} > 0.$$

Accordingly, an increase in the confidence in the subjective signal (γ_s) leads to an increase in its weight in the admissions process. In particular, the marginal effect is largest when γ_s is close to zero or one.

3.3. **Graduate Program and Job Placement.** Krueger and Wu (2000) observe that the relationship between objective GRE scores and outcomes is weaker if one conditions upon those individuals who are admitted to a program. Notice that since we have assumed the success function to be increasing and differentiable, $\lim_{\hat{\alpha}\to-\infty} p(\hat{\alpha}) = \underline{p} \geq 0$ and $\lim_{\hat{\alpha}\to\infty} p(\hat{\alpha}) = \bar{p} \leq 1$ exist and the success function is S-shaped, as illustrated in figure 2. The pool of applicants is assumed to have skill $\alpha_i \sim N(0, 1)$. After selection, we can suppose (for simplicity) that the pool of accepted students has a distribution of ability $\alpha_i \sim N\left(\bar{\alpha}, \bar{\sigma}^2\right)$, where mean ability satisfies $\bar{\alpha} > 0$ and variance is lower, $\bar{\sigma}^2 < 1$.

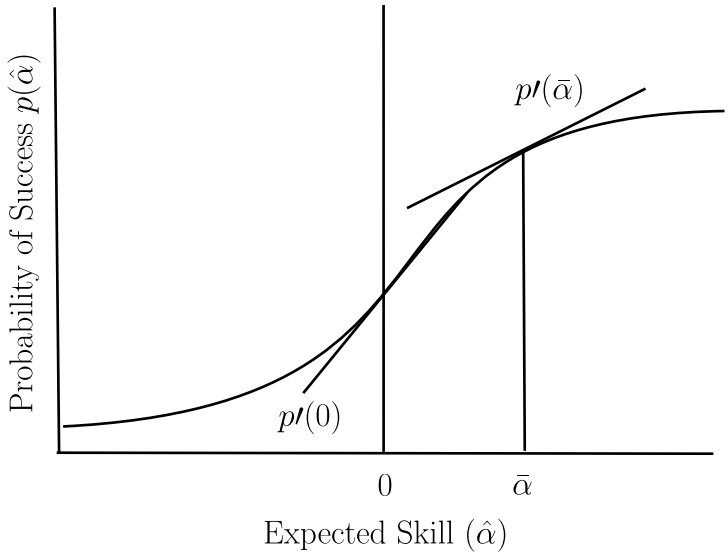


FIGURE 2. Success Function $(p(\hat{\alpha}))$

We can approximate $p(\hat{\alpha}_i)$ for each case $\hat{\alpha} \in \{0, \bar{\alpha}\}$ by a Taylor series expansion:

$$p(\alpha_i) \simeq p(\hat{\alpha}) + (\alpha_i - \hat{\alpha}) p\prime(\hat{\alpha}). \tag{9}$$

As we can see in figure 2, an implication of the higher mean scores is that in this figure $p\prime(\bar{\alpha}) < p\prime(0)$, leading to a decrease in the sensitivity of expected success with respect to skill. We can use this approximation to illustrate how the relationship between test scores and success varies with the subjective score. Suppose one uses a linear probability model to estimate the effect of the objective score on admissions:

$$a_i = \alpha_i + \beta g_i + \epsilon_i,$$

where $a_i = 1$ if person i is admitted, and zero otherwise. Using (4) and (9) we get:

$$\beta = \frac{cov(y_i, g_i)}{var(g_i)} = \frac{\gamma_g var(\alpha_i)}{\gamma_g^2 var(\alpha_i) + \left(1 - \gamma_g^2\right)} y'(\bar{\alpha}) \tag{10}$$

For the initial sample, $\alpha_i \sim N(0,1)$ one has $\beta^0 = \gamma^g p\prime(0)$. Post-selection one has:

$$\beta^S = \gamma_g \frac{p\prime(\bar{\alpha})}{\gamma_g^2 + \left(1 - \gamma_g^2\right)/\bar{\sigma}^2}.$$

Notice that under the assumption we are in the concave part of the S-curve $(y\prime\prime(\bar{\alpha}) < 0)$ then $\partial\beta^s/\partial\bar{\alpha} < 0$ while $\partial\beta^s/\partial\bar{\sigma}^2 > 0$ and hence $\beta^S < \beta^0$. Thus, the estimated relationship between the objective score and success should fall for more highly selected samples, a prediction that we can test with this data. This theoretical result is consistent with Krueger and Wu (2000) and Athey et al. (2007) who find that, in a sample of admitted economics graduate students, the GRE score does not predict future job market success. Kuncel et al. (2010) review the large literature on the validity of GRE scores and find that it is predictive of future performance for individuals who apply to graduate school. These two results are consistent with each other since the evidence in Kuncel

et al. (2010) is based upon a larger, more representative sample than the ones considered Krueger and Wu (2000) and Athey et al. (2007).

3.4. **Additional Implications of Agency Theory.** References are not compensated, nor is there any formal evaluation of their information. From a technical perspective, this situation can be viewed as a "cheap-talk" game (Crawford and Sobel (1982)). For a reference with the sole goal of seeing their applicant admitted to a top program, it is in their interest to behave strategically and choose the highest possible subjective ratings for their applicants. As discussed in the Data and Descriptive Statistics section (4) below, references are explicitly asked to rate applicants relative to their peers. From the management literature, it is well known that there is a tendency to bias this information when there are no penalties for exaggerating an applicant's skills (Milkovich et al. (2017)).

We observe variation in ratings, and hence we can reject the hypothesis that all references top-rank their applicants. Additionally, the subjective and objective scores measure, in effect, the extent to which admissions committees believe that these signals are valid measures of performance. We find evidence that the subjective evaluations from references are used by admissions committees to make decisions regarding applicant quality. Taken together, this provides some evidence that references and admissions committees share some common goal in ensuring that applicants are well matched. As Farrell and Rabin (1996) point out, cheap talk games do not make unique predictions, and hence, we need some evidence to provide guidance on how parties create and interpret these signals.

The Spence (1973) signaling model is another influential information model. It makes the prediction that when signaling by an individual is costly, the intensity of the signal should vary with unobserved signal costs. In our setting, a natural measure of signaling cost is the length of a recommendation letter. We should expect that longer letters signal a more qualified applicant because longer letters take more time to write.

At the same time, different references are likely to bias their evaluations to different degrees. One way to de-bias evaluations is to use relative performance evaluation (Green and Stokey (1983)). In our data, many references evaluate more than one applicant and we can construct relative measures of subjective rating, letter length, and letter sentiment. If the theory of relative performance evaluation is correct and references are likely to vary systematically in their verbosity, then we would expect that a reference's relative ranking of applicants provides useful information in addition to the level ratings. We have two ways of evaluating this information. First, we can ask if the relative ratings are related to other measures, such as the objective score. This simply asks if there is information in this signal. Second, we ask if the relative performance evaluation has an effect upon admissions. From expression (7) it follows that an increase in precision leads to more weight on that signal. Thus, if a reference writes more than one letter, the implication is that, via the relative performance signal, the weight placed on their views should be higher. This may lead to higher admissions rates by these references; a hypothesis we can test with this data.

3.5. **Applicant Incentives.** An optimal admissions policy implies that there is a cutoff level, $\hat{\alpha}^*$, satisfying $p(\hat{\alpha}^*) = \lambda^*$, where λ^* is the optimal Lagrange multiplier from proposition (1). Given that applying to programs is expensive, we would expect applicants to take into account the probability

of success when deciding whether to apply to a program or not. The belief of individual i can be represented by two parameters: the expected skill $\hat{\alpha}_i$ and the uncertainty that they have in this expectation, given by variance $\sigma_i^2 > 0$. Thus, the applicant supposes that the admissions committee observes a signal regarding their skill given by:

$$s_i = \hat{\alpha}_i + \sigma_i \epsilon_i,$$

where $\epsilon_i \sim N(0,1)$. Observe that decreasing uncertainty is equivalent to increased confidence in their self-assessment, which in turn leads to a more precise signal. Let individual $i's$ cost of making an application to a graduate program be c_i, and let the payoff if the application is successful be u_i. Person i will apply if and only if:

$$\Pr[a_i = 1] \times u_i - c_i > 0, \tag{11}$$

where $a_i = 1$ if they are accepted, and zero otherwise.

Clearly, an increase in desire to attend a graduate program (higher u_i), or lower application costs (lower c_i) increases the rate of application, conditional on the probability of acceptance. A less obvious result concerns the effect of beliefs on the probability of acceptance. Suppose that the cutoff for admission, given by $\hat{\alpha}^*$, where $y(\hat{\alpha}^*) = \lambda^*$, is known. A person's perceived probability of success is given by:

$$\Pr[a_i = 1] = \Pr[s_i \geq \hat{\alpha}^*] = F\left(\frac{(\hat{\alpha}_i - \hat{\alpha}^*)}{\sigma_i}\right),$$

where $F(\cdot)$ is the standard normal cumulative probability distribution. Given this, we can start with a pool of applicants with beliefs $\hat{\alpha}_i \sim N\left(m, \sigma_m^2\right)$. Clearly, as the criteria for admission rises, the number of applicants will fall. In particular, applicants observe their GRE scores before applying to graduate school. Thus, we would expect applicants with lower GRE scores to self-select out of the pool of applicants. An interesting implication of this selection is the effect of uncertainty (or lack of confidence):

$$\frac{\partial \Pr[a_i = 1]}{\partial \sigma_i} > 0 \text{ iff } \hat{\alpha}_i < \hat{\alpha}^*.$$

This expression illustrates the potentially complex relationship between self-assessment and applicant behavior. One way uncertainty is increased in practice is occurs when schools drop the requirement that individuals submit standardized test scores. In particular, if a program drops the requirement of submitting a GRE score, then this decreases the precision of the applicant's signal observed by the admissions committee and hence should increase σ_i. Applicants who believe their skill level is below the cutoff required for admissions with the GRE requirement may conclude that their chances of admissions have increased in the absence of a GRE score requirement. However, individuals who believe that their skill is above the cutoff with the GRE score may not apply and may choose to apply to a more competitive program.

Notice that the predicted effect of removing the standardized test score requirement is indeterminate—it depends upon the relative number of individuals who change their choices above and below the admission cutoff score. For highly selective programs, the number of individuals who believe they are above the cutoff is likely to be much less than the pool of individuals who

believe they do not meet the cutoff. Thus, it is not surprising that when Harvard dropped the SAT requirement due to the covid-19 pandemic, the number of applications increased (Lu and Tsotsong (2021))[9]

These observations are also consistent with a number of studies. Hoxby and Avery (2013) point out that knowledge and expectations play an important role in admissions. Goodman (2016); Goodman et al. (2020) show that knowledge about one's SAT scores has a major impact on admissions, confirming the role that self-assessed potential $\hat{\alpha}_i$ plays in application strategy. Similarly, Mulhern (2020) has found that the acceptance probability based upon test scores has an impact on application behavior.

Finally, the fact that success depends on application behavior highlights the point that the causal effect of evaluation is not merely a statistical question. Test requirements influence individuals' choices to invest in test preparation services and materials, which in turn affect admissions success and social stratification (see MacLeod and Urquiola (2015)). In this paper, we use admission to a top-ten economics program as a measure of success. But we do not observe application strategies. Thus, our success measure is necessarily a composite of application strategy and applicant information. Thus, like most employers, admissions committees can only evaluate those individuals who apply based upon the information submitted in an application, not upon their application strategy.

4. Data and Descriptive Statistics

The application process for PhD programs is highly structured and similar across programs. In most cases, any student with an undergraduate degree can apply after paying a fee. Our dataset consists of 6,320 applications from 2013 to 2019 for a single program. The application dataset consists of undergraduate transcripts, demographic information, a resume, essays, GRE scores, English proficiency scores, and evaluations provided by at least three references. Students with missing GRE scores are dropped.

Our goal is to assess the relative importance of measures that are measured systematically and in the same way for each applicant. As such, we do not use students' grade point average as a measure. There are more than 1,000 universities in our sample, and these universities exhibit widely different methods for measuring coursework performance. Thus, we use the three GRE test score percentiles from the quantitative, verbal, and writing sections, the subjective ratings and the text of the letters from the references.

The GRE is a standardized test that is an admissions requirement for many graduate schools in the United States and abroad. The exam consists of three sections: verbal reasoning, quantitative reasoning, and analytical writing. The GRE is administered by the Educational Testing Service (ETS) irrespective of an applicant's country of origin or undergraduate institution. ETS scores the verbal and quantitative sections on the same raw numerical scale, while the writing section is scored on an independent scale. ETS provides percentiles corresponding to the raw scores in each of the

9

See also https://www.nytimes.com/live/2021/01/25/world/covid-19-coronavirus/more-students-are-applying-to-elite-universities-after-test-scores-became-optional?smid=url-share.

three sections. For consistency in our analysis, we use these percentiles instead of the raw numerical scores .

Since many applicants attend higher-ranked programs, we refrain from simply using the admissions decision of the program providing this data as the main outcome variable. Crucially for the analysis, we obtain publicly available records of eventual admissions or job outcomes. These records include a combination of student directories from graduate program websites, LinkedIn profiles, personal websites, and news articles. While many applicants go on to attend an economics PhD program, some attend graduate degree programs in other fields, and some do not attend an academic program at all. Table 1 reports the fraction of applicants that we observe attending a top-ten economics PhD program. We assume that attending a top-ten program approximates at least one offer of admission to a top-ten program.[10] This outcome can be viewed as aggregating the views of the ten admissions committees, and hence provides a more general view of the role of evaluations in admissions.

Overall, from our pool of applicants, there is a 20.7% chance of admission to a top-ten program. Under the hypothesis that admissions are random for our pool, we can compute a confidence interval on the number of individuals who would be admitted. The 90% confidence interval is from 1,256 to 1,362 individuals. The largest group (68.7%) is the set of male applicants. For men, the probability of admissions is higher than the mean, at 21.8%. Women comprise 31.3% of the applicant pool, of whom 18.2% are admitted to a top-ten program. As a group, women are less likely to be admitted than men.

The work of Goldin and Rouse (2000) established that the perception of fairness or differential treatment may lead to a larger applicant pool that is of lower quality. Goldin and Rouse (2000) found that with blind, gender-neutral auditions there was a slight decrease in the average quality of the female musicians applying for positions. When we decompose the female group into US and non-US applicants, we observe positive selection for US female applicants. They form only 8.2% of the pool of applications, but 28.0% of this pool is admitted. In contrast to US female applicants, the admission rate for non-US female applicants, 14.8%, falls below the lower bound for the 90% confidence interval. We address the question of whether this is due to differential treatment or can be explained by characteristics in A.8 on identity and admissions.[11]

Table 1 also reports summary statistics of GRE percentile scores in our sample. The mean quantitative GRE is very high, and the mean of the accepted students is, as one would expect, higher. The verbal and writing scores are lower, consistent with the hypothesis that students self-select based upon their quantitative GRE scores.

[10] See A.5 for the definition of top-ten program used in this paper. Note that we observe individuals who attend top-ten programs and are not in our data. This implies that our data does not include everyone who applies to a top-ten program. As we discuss in the previous section, our goal is to assess the relative importance of signals for our sample.

[11] We do not include results for underrepresented minorities in this study for a number of reasons. First, all applicants respond to the sex and citizenship questions, and hence this data is comprehensive. Race is a self-reported question for US applicants only. Moreover, the number of applicants from an underrepresented minority group is very small (187 in our sample), and hence these estimates are likely to be noisy.

We consider the relative effects for the four identity-based groups listed in table 1. Non-US males make up close to half of the pool of applicants, followed by non-US females. Both groups have higher mean quantitative GRE scores than their corresponding US groups. However, the US applicants are admitted at a higher rate, and the quantitative GRE score gap is smaller in the admitted group. As one might expect, the US applicants have higher scores for the verbal and writing GREs.

TABLE 1. Summary of Objective Test Scores

Group	Application Pool (6,320)				Accepted to Top 10 (1,309)			
	Pool (%)	Quant	Verbal	Writing	Admit Rate (%)	Quant	Verbal	Writing
All Applicants	100%	92.2	79.3	59.4	20.7%	95.1	86.1	69.1
		(8.6)	(19.5)	(27.5)		(3.9)	(15.4)	(25.4)
US Male	19.8%	91.6	90.2	78.5	26.0%	95.2	94.2	85.4
		(9.5)	(12.9)	(20.7)		(3.5)	(7.7)	(15.9)
US Female	7.4%	89.5	89.6	80.9	28.4%	93.9	95.2	86.2
		(10.2)	(12.9)	(18.7)		(5.9)	(5.8)	(15.5)
Non-US Male	48.9%	93.2	75.2	51.9	20.1%	95.3	81.2	59.6
		(6.8)	(19.9)	(26.4)		(3.3)	(16.9)	(25.3)
Non-US Female	23.9%	91.5	75.3	52.2	15.1%	94.9	82.5	61.6
		(10.1)	(20)	(25.6)		(4.4)	(16.1)	(24.5)
Ivy Plus	9.5%	93.8	92.2	81.8	44.4%	95.4	94.3	85.5
		(6.7)	(11.3)	(18.5)		(3.6)	(8.2)	(15.8)

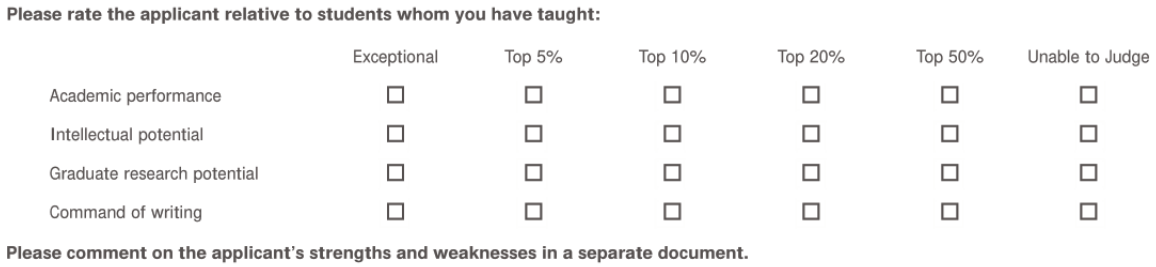
Note. Standard deviation in parentheses. Admit rate is the percent of that group who attend a top-ten program.

The subjective scores are constructed from the evaluations submitted by applicants' references. Each applicant is required to have at least 3 references. Our dataset includes 20,234 recommendation letters from a total of 10,800 references. Table 2 illustrates the distribution of references by the number of evaluations submitted. The majority of references provide only one evaluation, corresponding to 36.7% of the total number of evaluations. Individuals who evaluate five or more applicants comprise just 7.1% of the pool of references, yet account for 30.6% of the evaluations. Specifically, of the 10,800 distinct letter writers, 3,383 wrote two or more recommendation letters over the sample period.

TABLE 2. Number of Evaluations per Reference

Number of Letters Written	**Number of References**	**Percent of References**	**Number of Letters**	**Percent of Letters**
1	7,444	68.8%	7,444	36.7%
2	1,586	14.6%	3,172	15.6%
3	653	6.0%	1,959	9.7%
4	375	3.5%	1,500	7.4%
5+	769	7.1%	6,203	30.6%
2+	3,383	31.2%	12,834	63.3%
Total	10,827		20,278	

The evaluations consist of a letter of recommendation and a subjective rating of the applicant. As depicted in figure 3, the reference is asked to check a box for each dimension: academic performance, research potential, intellectual potential, and writing skill. These scores range from "Exceptional", coded as 1, to "Top 50%", coded as 5, with "Top 5%", "Top 10%", and "Top 20%" in between. Letter writers can also mark "Unable to Judge" or choose to leave the rating blank. Figure 4 illustrates a histogram of the subjective ratings. The lower dark portion of the bars illustrates the fraction admitted to a top-ten program. The subjective rating provides a measure of skill relative to others in an evaluator's reference group. The graph shows that on average, students score highest on research potential and intellectual potential while on average the score is lower for writing ability. Figure 4 exhibits a well-known feature of subjective scores: they are highly compressed at the top (see Milkovich et al. (2017)). Over half of evaluations rate an applicant in the top 5% of their reference group.



Please rate the applicant relative to students whom you have taught:

	Exceptional	Top 5%	Top 10%	Top 20%	Top 50%	Unable to Judge
Academic performance	☐	☐	☐	☐	☐	☐
Intellectual potential	☐	☐	☐	☐	☐	☐
Graduate research potential	☐	☐	☐	☐	☐	☐
Command of writing	☐	☐	☐	☐	☐	☐

Please comment on the applicant's strengths and weaknesses in a separate document.

FIGURE 3. Subjective Questions Form

In addition to the subjective ratings and objective test scores provided in the application package, we process 20,234 recommendation letters into analysis-ready data, using optical character recognition (OCR) techniques to extract text from letters that were submitted as images. As is standard in textual analysis, we remove stop words such as articles and conjunctions (Gentzkow et al. (2019)). After removing stop words, the average recommendation letter length is 355 words. We follow the previous literature on analyses of academic recommendation letters (Trix and Psenka (2003), Madera et al. (2009), and Dutt et al. (2016)) and use seven topic dictionaries as letter features: standout, ability, research, grindstone, teaching, communal, and agentic. For example, grindstone words denote diligence and work ethic, while communal words capture interpersonal traits and general amicability.[12] In addition to these topic frequencies, we include measures of positive and negative sentiment word patterns from the Lexicoder Sentiment Dictionary (LSD) of Young and Soroka (2012). We standardize these features by dividing the dictionary word count by the letter length. We average the letter features across all letters in an application, reporting the average and median from the application level. Table 3 shows that, on average, teaching terms are the most frequent in letters, followed by research, then grindstone terms. Meanwhile, the dimensions that emphasize an applicant's personality or demeanor—communal and agentic—are the least frequent.

[12]See appendix section A.6 for complete definitions of the letter feature dictionaries.

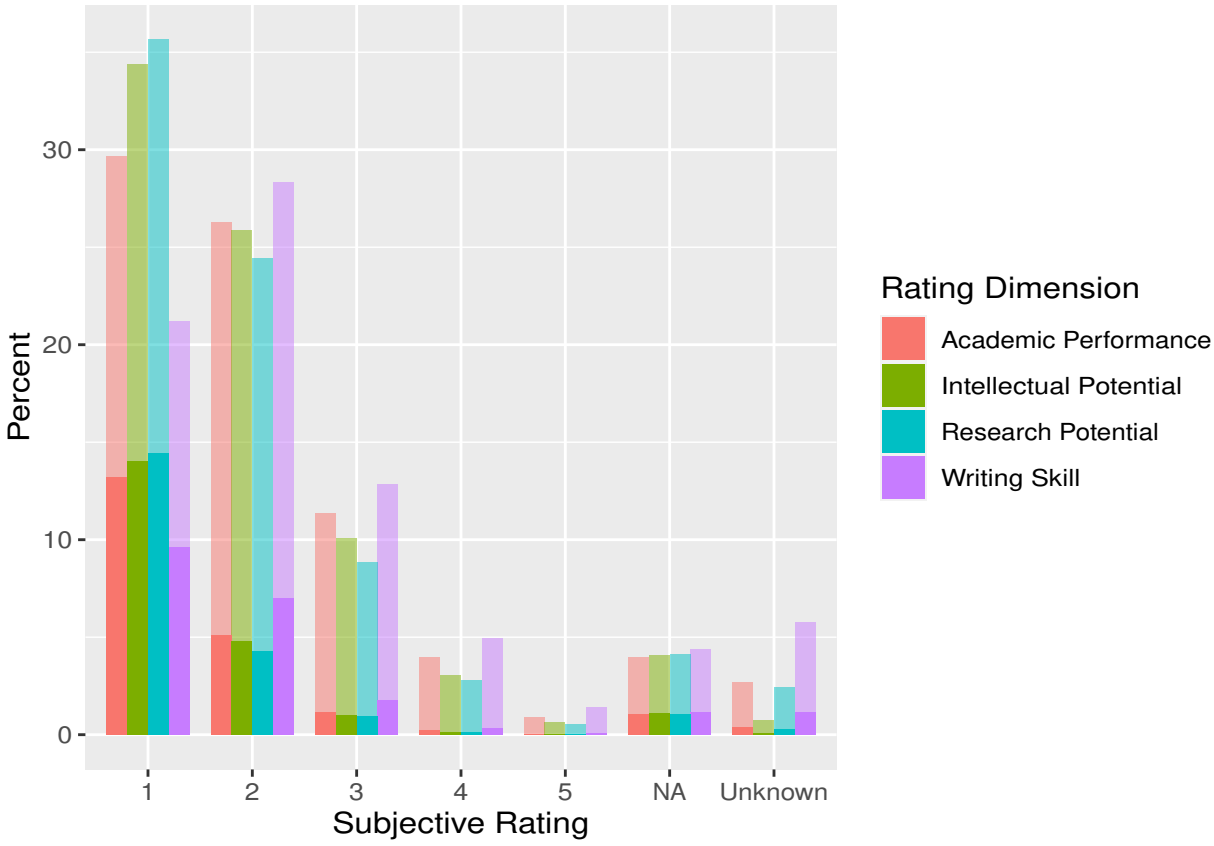


FIGURE 4. Subjective Rating Distribution

TABLE 3. Letter Features by Subgroup

Panel B: Letters Features

		Applied		Accepted	
Group	Score Category	Mean	Median	Mean	Median
Pooled	Length	354.4 (124.9)	335.3	412.6 (126.7)	393.7
	Standout Terms (%)	0.6 (0.4)	0.6	0.7 (0.3)	0.7
	Ability Terms (%)	1.2 (0.5)	1.2	1.1 (0.4)	1.1
	Research Terms (%)	3.2 (1.3)	3.1	3.4 (1.2)	3.3
	Grindstone Terms (%)	1.5 (0.7)	1.5	1.5 (0.6)	1.4
	Teaching Terms (%)	3.4 (1.3)	3.4	3.3 (1.3)	3.3
	Communal Terms (%)	0.5 (0.3)	0.4	0.4 (0.3)	0.4
	Agentic Terms (%)	0.2 (0.2)	0.2	0.2 (0.2)	0.2
	Positive (%)	9.1 (1.7)	9.0	8.9 (1.5)	8.8
	Negative (%)	1.8 (0.7)	1.8	1.9 (0.7)	1.8
US Males	Length	387.6 (122.3)	374	438.3 (121.6)	428.3
	Standout Terms (%)	0.6 (0.3)	0.6	0.7 (0.3)	0.6
	Ability Terms (%)	1.3 (0.5)	1.3	1.2 (0.5)	1.2
	Research Terms (%)	3.5 (1.4)	3.5	3.8 (1.3)	3.8
	Grindstone Terms (%)	1.7 (0.6)	1.7	1.7 (0.6)	1.7
	Teaching Terms (%)	3.1 (1.4)	3.0	2.8 (1.2)	2.7
	Communal Terms (%)	0.4 (0.3)	0.4	0.3 (0.2)	0.3
	Agentic Terms (%)	0.2 (0.2)	0.2	0.2 (0.1)	0.2
	Positive (%)	9.3 (1.6)	9.2	9 (1.4)	9.0
	Negative (%)	2.0 (0.7)	1.9	2.1 (0.7)	1.9
US Females	Length	403.6 (124.9)	391	463.8 (120.6)	461.2
	Standout Terms (%)	0.6 (0.3)	0.6	0.6 (0.3)	0.6
	Ability Terms (%)	1.3 (0.5)	1.3	1.2 (0.4)	1.2
	Research Terms (%)	3.6 (1.2)	3.6	4 (1.1)	4.0
	Grindstone Terms (%)	1.8 (0.7)	1.8	1.9 (0.6)	1.9
	Teaching Terms (%)	3 (1.3)	2.8	2.7 (1.2)	2.4
	Communal Terms (%)	0.4 (0.3)	0.4	0.4 (0.2)	0.4
	Agentic Terms (%)	0.2 (0.2)	0.2	0.2 (0.1)	0.2
	Positive (%)	9.4 (1.6)	9.3	9.1 (1.3)	9.3
	Negative (%)	2.0 (0.7)	1.9	2.1 (0.7)	2.1
Non-US Males	Length	345.1 (124.9)	327	394.1 (124.2)	376.8
	Standout Terms (%)	0.6 (0.4)	0.6	0.7 (0.4)	0.7
	Ability Terms (%)	1.2 (0.5)	1.1	1.1 (0.4)	1.1
	Research Terms (%)	3.1 (1.2)	3.0	3.1 (1.1)	3.0
	Grindstone Terms (%)	1.4 (0.6)	1.3	1.3 (0.6)	1.2
	Teaching Terms (%)	3.6 (1.2)	3.6	3.7 (1.2)	3.6
	Communal Terms (%)	0.5 (0.3)	0.4	0.4 (0.3)	0.4
	Agentic Terms (%)	0.2 (0.2)	0.2	0.2 (0.2)	0.2
	Positive (%)	8.9 (1.7)	8.8	8.8 (1.5)	8.7
	Negative (%)	1.8 (0.7)	1.7	1.8 (0.7)	1.7
Non-US Females	Length	331 (117.5)	310.8	396.4 (130.1)	364.5
	Standout Terms (%)	0.6 (0.4)	0.6	0.7 (0.3)	0.7
	Ability Terms (%)	1.2 (0.5)	1.1	1.1 (0.4)	1.1
	Research Terms (%)	3.0 (1.2)	2.9	3.2 (1.1)	3.2
	Grindstone Terms (%)	1.5 (0.7)	1.5	1.4 (0.6)	1.4
	Teaching Terms (%)	3.6 (1.3)	3.6	3.5 (1.2)	3.6
	Communal Terms (%)	0.5 (0.3)	0.5	0.4 (0.3)	0.4
	Agentic Terms (%)	0.2 (0.2)	0.2	0.2 (0.2)	0.2
	Positive (%)	9.3 (1.9)	9.1	8.9 (1.6)	8.8
	Negative (%)	1.8 (0.7)	1.7	1.9 (0.7)	1.8

5. Construction of the Performance Measure Scores

In this section, we outline the construction of subjective and objective performance measure scores. In order to discern the relative importance of objective versus subjective evaluations, we include measures of both in our base model, including variables discussed in the previous section as well as some dummy variables to capture non-linearities in the data. In total, we include 53 variables in the base model.

Taking that an applicant faces a fixed admissions probability of p^0 as the null hypothesis, we estimate a logit model to measure the extent to which admissions vary from random as a function of observed characteristics. This approach allows us to measure the relative importance of objective versus subjective evaluations.

Recent work by Jeganathan et al. (2021) shows that from all the possible algorithms one could use, the logit model is the most appropriate for admissions data. The L1 regularized (LASSO) logit model with cross-validation is the basic machine learning model (Hastie et al. (2009)) that performs well for dichotomous outcome variables. Waters and Miikkulainen (2014) also use this approach to build an algorithm to model and aid admissions to a computer science PhD program. Based on this work, we employ a logit model to aggregate observed characteristics. However, using a probit or linear probability model does not change the qualitative results for the relative importance of objective and subjective scores.

Let $\vec{x}_i \in X$ denote the observed characteristics of applicant $i \in I$, where the number of individuals is given by n. We estimate a model of the form:

$$Prob\left[a_i = 1\right] = F\left(\gamma + \beta\vec{x}_i\right), i \in I \tag{12}$$

where $F(\cdot)$ is the cumulative distribution function of the logit model, a_i is equal to one if person i is admitted to a top-ten program and zero otherwise. We use the full sample of applications to estimate the model to generate estimates $\hat{\gamma}$ and $\hat{\beta}$. From these we can define the predicted probability of success of individual i as:

$$p_i = F\left(\hat{\gamma} + \hat{\beta}\vec{x}_i\right), i \in I.$$

The score $\gamma + \beta\vec{x}_i$ corresponds to $\hat{\alpha}_i$ in the theory, while the predicted probability p_i corresponds to $y\left(\hat{\alpha}_i\right)$ in the theory. Notice that we can use p_i to rank applicants from the most skilled (high p_i) to the least skilled. This motivates the following random variable that can be viewed as a representation of a_i:

$$\delta_i = \begin{cases} 1, & with\, Prob\, p_i, \\ 0, & with\, Prob\, 1 - p_i. \end{cases}$$

Thus, δ_i is a prediction of whether a person is admitted to a top-ten program or not, and we define the quality of these predictions by their *node purity*. The node purity is a monotonic transformation of the likelihood function that lies in the interval [0,1]. This is shown in Appendix A.4. Node purity is used often in the machine learning literature, and the unit interval makes interpreting the value easier than interpreting the value of the likelihood. Throughout, we report the node purity in our logistic regression tables rather than the value of the likelihood.

First, we estimate model (12) to provide a benchmark measure of the relationship between different performance measures and admissions. This model allows us to explore the predictions of agency theory, the effect of identity, and subsequent job market success.

Our goal is not to provide the best empirical model of admissions, but rather to assess the relative information content of objective and subjective evaluations viewed from the perspective of the admissions committee.

Tables 20-22 in the appendix report the results from a two-step procedure to predict admissions. In the first stage, we run LASSO regressions with five-fold cross-validation to determine which variables best predict admissions. The analysis in the appendix A.7 shows that the selected model has good out-of-sample properties and that, even without regularization, the model has a very good out-of-sample fit. This is consistent with the hypothesis that the applicants can be viewed as independent draws from a sample, combined with the hypothesis that admissions committees evaluate each applicant on their merits. As such, the model provides a good representation of the effect of application characteristics, $\vec{x}_i$, on admissions. This procedure determines the variables to include in the model for score creation.

Table 20 reports the results for the GRE variables. The first column is the result from a logit that uses all the variables, normalized to mean zero and variance 1. We use both the score itself as well as the score quartile indicators to capture the non-linear effect of the GRE scores. The results from the LASSO with five-fold cross-validation are reported in the second column. Notice that all variables of the quantitative GRE score are retained. In particular, the level effect is 0.534 and is much larger than any of the other objective score variables. The fact that each quartile has a significant, but different, effect suggests that a simple cutoff rule is not being used. Column 3 uses only the LASSO-selected variables, and the variables in column 3 are not normalized. Comparing column 3 with column 1, we observe that all the significant variables in column 1 are retained by LASSO, and only the insignificant ones are dropped—this suggests LASSO performs well in selecting variables that contain information for top ten admissions. These effects also confirm the belief that the quantitative GRE test score plays an important role in admissions. Though, there is evidence that all GRE test scores play a role in admissions.

Next, we use weights in column 3 to construct the aggregate scores. We do not normalize the variables in column 3 because the aggregate scores we construct will be normalized to mean 0 and standard deviation 1. Normalizing the variables in column 3 and then normalizing them again when creating the aggregate scores could introduce noise to the score aggregation procedure. We construct the aggregate scores in the following way: using the GRE quantitative score as an example, we let $\vec{x}_i^{greq}$ be the vector of GRE quantitative variables (GRE Quant Score, GRE Quant Score = 2nd Quartile, GRE Quant Score = 3rd Quartile, GRE Quant Score = Top (4th) Quartile) and let $\hat{\beta}^{greq}$ be the corresponding estimated parameter values from column 3 in Table 20.

We define the *normalized GRE quantitative score* by:

$$\theta_i^{greq} = \frac{\hat{\beta}^{greq}\vec{x}_i^{greq} - mean_{i\in I^0}\left(\hat{\beta}^{greq}\vec{x}_i^{greq}\right)}{sd_{i\in I^0}\left(\hat{\beta}^{greq}\vec{x}_i^{greq}\right)}. \tag{13}$$

Thus, we z-score the aggregate score to have zero mean and variance of 1. Here we follow the common practice in the education literature to assess the effect of a test score in terms of standard deviations of the measure. Given that the variance of the score in the population is fixed to 1, we can run a regression to see how sensitive admissions is to a normalized score. We construct a score for each of the three GRE test scores, denoted by θ^{greq},θ^{grev} and θ^{grew} for quantitative, verbal, and writing, respectively.

Next, we follow this procedure for the subjective ratings provided by the references. The variables correspond to the four dimensions on the subjective rating form illustrated in figure 3. When a reference leaves a response blank, we record it as "NA" and use it as the baseline for indicator variable comparison. The subjective rating variables are the fraction of references giving that rating. For example, Academic Performance=1 is the fraction of that applicant's references who rate the applicant as "Exceptional" along the dimension of academic performance. Following the procedure given by equation (13), we use these variables to construct the subjective score θ_i^s.

Table 21 illustrates the factors that have an effect upon admissions. Interestingly, only a handful of these ratings have a positive effect, namely academic performance=1 and writing skill=1. The results of the subjective ratings have either no effect or a negative effect. Interestingly, a reference reporting that they do not have enough information is a negative signal in all cases except for writing skill. We use all the weights in column 3 of 21 to construct the subjective ratings score θ^s, following the same procedure as equation (13).

The next set of signals comes from the letters of recommendation, as shown in Table 22. This includes letter features like length and the frequency of certain words, as discussed in the data section. Letter length as measured by the number of words has a positive effect, and the number of pages in a letter has a weak positive effect when the number of pages is small, but a negative effect when the number of pages is large. In terms of letter language, standout words have a strong positive effect. Research words and negative words also have a weak positive effect, while ability, communal, and positive words have a negative effect. The fact that negative words have a positive weight while positive words have a negative effect on admissions suggests that admissions committees down-weight flattery and up-weight objectivity. Given that these metrics are normalized to the letter length, this result is consistent with the possibility that positive, communal, and ability words crowd out words on the performance of the individual. The main implication is that these simple measures of word types used in letters are associated with significant effects upon the probability of admission. Finally, we use the weights on Letter Length, page count, page count (1, 2], page count (2, 3], page count (3, 4] and page count 4+ in column 3 to create the letter length score θ^l, following similar steps in equation (13). Similarly, we use all the remaining weights in column 3 of 22 to create the letter text features score θ^w.

Table 4 summarizes the scores we have constructed. The GRE test scores form the basis of the objective summary score we construct, while the remaining scores based upon the data from the references are the basis of the subjective score. We include three relative scores based upon the subjective ratings, letter length, and letter text. As discussed above, individual references may have systematic biases. When a reference writes two or more letters, then this bias can be reduced by

using the comparative score (Green and Stokey (1983)). These scores are constructed as follows. Let $I(j)$ be the set of individuals who have a letter from reference j. Let $J(i)$ be the set of references for individual i. For each reference, we construct the mean score for $k \in \{s, l, w\}$ (corresponding to the subjective rating, length and letter text respectively) for reference j:

$$\bar{\theta}_j^k = \frac{1}{\#(I(j))} \sum_{i \in I(j)} \theta^k.$$

The corresponding relative score $k_r \in \{s_r, l_r, w_r\}$ for individual i is given by:

$$\theta_i^{k_r} = z - score \left(\frac{1}{\#(J(i))} \sum_{j \in J(i)} \left(\theta_{ij}^k - \bar{\theta}_j^k \right) \right).$$

If the reference has only one student then by construction $\theta_{ij}^k - \bar{\theta}_j^{k_R} = 0$, and this term will have no effect. In particular, if all the references for a student write only one letter, then $\theta_i^{k_r} = 0$.

TABLE 4. Individual Performance Measures

Variable	Name of score	Content
θ^{greq}	GREQ	GRE Quantitative Reasoning
θ^{grev}	GREV	GRE Verbal Reasoning
θ^{grew}	GREW	GRE Analytical Writing
θ^s	Subjective	Average of reference subjective ratings
θ^l	Length	Average letter length
θ^w	Writing	Average letter language features
θ^{s_r}	Relative subjective rating	Deviation from average subjective conditional on reference
θ^{l_r}	Relative length	Deviation from average length conditional on reference
θ^{w_r}	Relative writing	Deviation from average writing conditional on reference

Variables that are used to construct these scores: 1) θ^{greq}: GRE Quant Score, GRE Quant Score = 2nd Quartile, GRE Quant Score = 3rd Quartile, GRE Quant Score = Top (4th) Quartile; 2)θ^{grev}: GRE Verbal Score, GRE Verbal Score = 2nd Quartile, GRE Verbal Score = 3rd Quartile, GRE Verbal Score = Top (4th) Quartile; 3) θ^{grew}: GRE Writing Score, GRE Writing Score = 2nd Quartile, GRE Writing Score = 3rd Quartile, GRE Writing Score = Top (4th) Quartile; 4) θ^s: All the academic performance indicators, intellectual potential indicators, research potential indicators, and writing skill indicators. 5) θ^l: Letter Length, page count, page count (1, 2], page count (2, 3], page count (3, 4] and page count 4+; 6) θ^w: Ability Words, Standout Words, Research Words, Grindstone Words, Teaching Words, Communal Words, Agentic Words, Negative Words,Positive Words. The weights of aggregating the variables into scores are in column 3 of Tables 20, 21, and 22 in the appendix LASSO section. All aggregated scores are scaled to mean 0 and variance 1. The relative scores are created using the corresponding aggregated scores.

The next table documents the correlation between the scores. Notice that all the correlations are positive and, in particular, the quantitative GRE score is positively correlated with all the other measures. This is consistent with these measures having a common α_i component. The imperfect correlation can be due to two factors. The first, as illustrated in the theory section, is that they are noisy measures of the same underlying component. The differences in correlation are due to variation in the variance of the noise component. The second possibility is that they are measuring a different component of performance, such as non-cognitive skills like enthusiasm or desire to study economics, that may not be captured in the GRE score. Furthermore, the subjective rating, θ^s, is

more highly correlated with the writing GRE score, θ^{grew}, than with the GRE quantitative score, θ^{greq}. However, we see the reverse with the text score for letters, θ^{w}.

Also, the relative writing score and relative letter length both have a high correlation with the writing and letter scores. Given that the relative score for references that write a single letter is zero, this suggests there is a great deal of variation in a reference's language between the students for whom they are writing a letter. If the correlation were due to the common zeros, then we would expect the correlation between relative scores to be similar to these, which they are not. The next section assesses the importance of these scores for admissions.

TABLE 5. Covariance of Performance Measures

	θ^{greq}	θ^{grev}	θ^{grew}	θ^{s}	θ^{l}	θ^{w}	θ^{s_r}	θ^{l_r}	θ^{w_r}
θ^{greq}	1								
θ^{grev}	0.152	1							
θ^{grew}	0.049	0.537	1						
θ^{s}	0.159	0.174	0.202	1					
θ^{l}	0.147	0.182	0.203	0.277	1				
θ^{w}	0.199	0.066	0.033	0.193	0.194	1			
θ^{s_r}	0.112	0.149	0.171	0.641	0.190	0.155	1		
θ^{l_r}	0.047	0.057	0.078	0.210	0.428	0.108	0.302	1	
θ^{w_r}	0.057	0.048	0.037	0.148	0.077	0.521	0.219	0.145	1

6. The Relationship Between Performance Measures and Admissions

The score measures defined in section 5 transform the complex, non-linear measures of performance into a vector of measures with mean zero and unit variance. In this section, we begin by assessing the relative importance of these measures in predicting admissions, which in turn provides some direct evidence of the predictions of agency theory. The first prediction was that admissions are based upon all measures of performance that provide useful information. Second, relative performance measures, by screening out individual variation, provide additional information. References are not rewarded, nor punished, for their subjective responses, hence cheap-talk models predict that there is an incentive to shade their responses in favor of their applicants. Finally, writing a letter is costly, and hence the signaling model predicts that increasing their length may be a credible signal of quality.

Table 6 reports the results of a logit regression using the scores constructed in the previous section. Since the aggregates are constructed using the full estimated model, the fit of the model using the z-scored variables is identical to the model used to construct the weights for score aggregation.

TABLE 6. Effect of Performance Measures on Top 10 Admissions

	Dependent variable: T10 Admission				
		Actual Data			Placebo Test
	(1)	(2)	(3)	(4)	(5)
Subjective Rating θ^s			0.7348	0.6790	0.001
			(0.0515)	(0.0482)	(0.045)
GRE Quant Score θ^{greq}		0.7779		0.6406	0.008
		(0.0654)		(0.0813)	(0.032)
GRE Verbal Score θ^{grev}		0.2929		0.2318	-0.003
		(0.0506)		(0.0471)	(0.036)
GRE Writing Score θ^{grew}		0.3234		0.1850	0.002
		(0.0317)		(0.0323)	(0.037)
Letters Features θ^w			0.4044	0.3840	-0.002
			(0.0605)	(0.0559)	(0.039)
Letter Length θ^l			0.4512	0.3866	0.008
			(0.0484)	(0.0481)	(0.036)
Relative Subjective Rating θ^{s_r}			0.3733	0.3049	-0.001
			(0.0565)	(0.0559)	(0.043)
Relative Letter Features θ^{w_r}			−0.0597	−0.0622	-0.001
			(0.0540)	(0.0533)	(0.036)
Relative Length Score θ^{l_r}			−0.0559	−0.0309	-0.002
			(0.0541)	(0.0478)	(0.036)
Intercept	−1.3424	−1.5790	−1.7673	−1.9182	-1.345
	(0.0424)	(0.0568)	(0.0755)	(0.0871)	(0.001)
Pseudo $R^2 \in [0,1]$	0	0.0927	0.1756	0.2169	0.001 (0.001)
Node Purity	0.6004	0.6295	0.6567	0.6707	0.601 (0.0002)
Observations	6,320	6,320	6,320	6,320	6,320

Note. Standard errors are clustered by application year and are heteroskedasticity-robust (HC3 - pseudo-jackknife). For the placebo test, the standard errors are computed from Monte-Carlo simulation of point estimates. For details of HC3 - pseudo-jackknife standard errors, please see the paper by MacKinnon (2013).

Column 1 is the model with no explanatory variables. The intercept, −1.3424, represents the estimated probability that a randomly chosen person is admitted. This probability is given by $logit^{-1}(-1.3424) = 20.7\%$,consistent with table 3. The node purity, the monotonic transformation of the log likelihood value, is 0.6. It is greater than 0.5 because the admissions probability is less than 0.5. The pseudo-R^2 is a measure in $[0,1]$ of the improvement of fit due to the explanatory variables, and hence is normalized to zero when there is only an intercept term.

Column 2 reports the outcome from a logit model where the explanatory variables are the objective GRE scores. The value of the intercept falls significantly, consistent with the fact that GRE scores are correlated with admissions. The pseudoR^2 reports a 10% fit, consistent with the increase in node purity from 0.6 to 0.63. Interestingly, if we use only subjective scores (reported in column 3) then the fit is better, with a pseudo-R^2 reporting an 18% fit, suggesting that the subjective evaluations are more important for admissions than the objective GRE scores. In addition, the relative subjective rating has a significant effect, while the other relative scores do not.

We can get a better sense of both their individual contribution and relative importance by including all the variables in the regression, which is reported in column 4. The fit improves again with a pseudo R^2 of 22%. The subjective rating has the largest weight, followed by the GRE quantitative score, which has a similar weight. The final column performs a placebo test with this model. We randomly allocate seats, retaining an overall acceptance rate of 20.7% observed in the data. As one can see, all the coefficients are zero, and the pseudo R^2 is zero, consistent with the true model of these variables having no effect, and hence we can reject the hypotheses that admissions are random.

6.1. **Relative Performance Measures.** One benefit of objective test scores is that they are the result of a standardized test that allows for consistent comparisons across individuals. This benefit is missing for subjective evaluations. Agency theory proposes a potential solution when the same reference evaluates several applicants. The relative performance as measured by the subjective score should have a large and significant effect, while the other relative scores have little effect. This can be due to either the fact that the subjective rating is more informative or the rich narrative in a letter provides precise information. Another alternative is that human admissions committees simply cannot respond to this information since it requires summarizing subjective evaluations from thousands of applications.

If there is information in the relative scores, the relationship between relative scores and other scores from the same references will be significant and the models with relative scores will have a better fit. Table 7 reports the correlation between signals and measures of letter length. Column 1 shows that all performance levels are positively correlated with letter length, consistent with the hypothesis that this provides a positive signal of performance. However, we can see in column 2 that the letter length score is uncorrelated with the relative subjective rating and is negatively correlated with relative letter features. This is evidence that references vary the length of their letters; however, the longer letters appear to be padded with terms that reduce the overall effectiveness of the letter. Since the letter length is uncorrelated with the within-reference subjective score, this suggests that references are not systematically writing longer letters for individuals who they more highly value as measured by the subjective scores. Moreover, though they write longer letters, they appear to use words that are not associated with a higher admissions probability. As we can see from table 22, there are some counterintuitive effects that are consistent with this result. In particular, more "ability" and positive words have a negative effect upon admissions. Hence, given that talking about positive features of an applicant is a natural way to "pad" a letter, doing so might generate the negative within-reference effect.

The next two columns look at the relationship of scores with letter features. This measure is positively correlated with all performance measures except the writing score. In particular, it is positively correlated with the relative subjective rating, indicating that this score contains positive information. In the case of letter length, the correlation between letter length and subjective score is unaffected by the introduction of the relative scores, but this is not the case for letter features. Here we see a significant drop in this effect when we add the relative subjective rating, consistent with the hypothesis that the relative score contains useful information.

TABLE 7. Correlation of Signals and Letter Length

	Dependent Variable			
	Letter Length θ^l		Letter Features θ^w	
	(1)	(2)	(3)	(4)
Subjective Rating θ^s	0.2044	0.2012	0.1347	0.1038
	(0.0123)	(0.0153)	(0.0132)	(0.0168)
Letter Features θ^w	0.1323	0.1573		
	(0.0135)	(0.0159)		
Letter Length θ^l			0.1390	0.1350
			(0.0144)	(0.0161)
GRE Quant Score θ^{greq}	0.0723	0.0695	0.1574	0.1573
	(0.0125)	(0.0126)	(0.0144)	(0.0144)
GRE Verbal Score θ^{grev}	0.0618	0.0621	0.0133	0.0125
	(0.0142)	(0.0142)	(0.0146)	(0.0146)
GRE Writing Score θ^{greq}	0.1207	0.1207	−0.0368	−0.0383
	(0.0142)	(0.0142)	(0.0147)	(0.0147)
Relative Subjective Rating θ^{s_r}		0.0093		0.0476
		(0.0163)		(0.0160)
Relative Letter Features θ^{w_r}		−0.0486		
		(0.0143)		
Relative Length Score θ^{l_r}				0.0092
				(0.0134)
Constant	0.0000	0.0000	0.0000	0.0000
	(0.0118)	(0.0117)	(0.0121)	(0.0120)
Observations	6,320	6,320	6,320	6,320
R^2	0.1285	0.1302	0.0841	0.0856
Adjusted R^2	0.1279	0.1292	0.0833	0.0846

Note. Standard errors are heteroskedasticity-robust (HC3).

From these results, we conclude that there is information contained in the relative subjective rating. This does not imply that a change in the relative score is in fact perceived by the admissions committees. Rather, it shows that the impression committees form regarding an applicant is correlated with a reference's responses to the subjective questions. We find, as predicted by the Spence signaling model, that letter length is correlated with admissions. However, this does not seem to be a *within*-reference effect. The results on the relative length measure and relative letter feature measures are not consistent with a reference writing a longer letter to signal a better applicant. Rather, the negative effect of a text measure is consistent with the references writing a longer letter for applicants that they like! This follows from the negative effect of positive words. However, there is evidence of selection across references. Applicants who choose references that write longer letters on average will do better than those that do not. Below we provide a further exploration into the characteristics of references that may be generating these results.

6.2. **Subjective versus Objective Performance Measures.** In this section, we evaluate the importance of the objective GRE scores within the pool of applicants compared to the importance

of the subjective scores. This exercise comes with the caveat that the relative importance of objective and subjective scores depends upon both the behavior of the admissions committees and the application strategy of the applicants. In particular, individuals know and use GRE test scores to decide whether or not to apply to particular graduate programs, and individuals with lower GRE test scores apply less frequently to highly ranked programs. As such, our measure is likely an upper bound on the importance of GRE test scores because students with lower GRE scores may not apply to highly ranked programs, thereby lowering their probability of a top ten admission.

We follow the same procedure to construct the scores as in the previous section. We use the model estimated in column 3 of table 6 to weight the various performance measures. We define:

$$\theta^{obj} = z - score(\beta^{greq}\theta^{greq} + \beta^{grev}\theta^{grev} + \beta^{grew}\theta^{grew}),$$
$$\theta^{sub} = z - score(\beta^{s}\theta^{q} + \ldots + \beta^{w_r}\theta^{w_r}).$$

As before, the model fit for the full sample of applicants when using these variables as explanatory variables is identical to the fully specified model. The benefit of this procedure is that it allows us to compare the relative importance of objective scores against subjective scores. The correlation between these scores is reported in table 8. Since these scores have a mean of zero, the intercept is zero. We observe positive correlation between the two variables, and hence both provide useful information regarding an individual's suitability for graduate school.

TABLE 8. Correlation between Subjective and Objective Scores

	Subjective Score - θ^S
Objective Score - θ^o	0.284
	(0.012)
Constant	0.000
	(0.012)
Observations	6,320
Adjusted R^2	0.081

Note. Standard errors are heteroskedasticity-robust (HC3).

To assess the relative importance of these signals, we report a regression on top ten admissions in table 9, which corresponds to the signal equation (4). In this table, we z-score the aggregate objective and subjective scores in order to compare their overall effects on admissions. The importance of the subjective score is larger than the objective score. Column 2 is the effect of the objective score, while column 3 is the effect of the subjective score. Both of these effects are larger individually than when taken together due to their correlation. In column 4, the estimated subjective score weight is 42% larger than the objective weight.

We can translate these differences into probabilities, which we report in table 10. The first column is the probability of admissions, corresponding to the first column in table 9. Next, column 2 reports the increase in the probability of admissions resulting from a one standard deviation increase in the objective score. The first row of column 2 corresponds to increasing the objective score by one standard deviation using column 4 in table 9, fixing the subjective score at zero. However, we know

TABLE 9. Subjective vs Objective Scores in Top 10 Admissions

	T10			
	(1)	(2)	(3)	(4)
Objective Score - θ^o		1.023		0.789
		(0.071)		(0.084)
Subjective Score - θ^S			1.265	1.125
			(0.059)	(0.056)
Constant	−1.342	−1.587	−1.725	−1.879
	(0.042)	(0.055)	(0.072)	(0.084)
Pseudo R^2	0	0.0916	0.1666	0.2101
Node Purity	0.6004	0.6292	0.6537	0.6683
Observations	6,320	6,320	6,320	6,320

Note. Standard errors are clustered by application year and are heteroskedasticity-robust (HC3).

that these scores are correlated, so an observed increase in the objective score is associated with an increase in the subjective score. This explains why the coefficient for the objective score in column 2 of table 9 is greater than in column (4), where there is a control for the effect of the subjective score. One standard deviation increase in objective score, while holding the subjective score constant is associated with a 21% increase in the admissions probability. However, an applicant in our pool whose objective score is one standard deviation higher than the average objective score has a 75% greater chance of admissions.

The third column reports the corresponding results for a one standard deviation increase in the subjective score. In the case with correlation, if an applicant has a one standard deviation subjective score increase relative to the mean score, this results in an 87% increase in admissions probability. The uncorrelated score increase is of independent interest if one assumes that the results are causal. That is, given an "average" applicant, if the references as a whole increase their evaluation of the applicant by one standard deviation, while holding the objective score fixed, corresponds to a 50% increase in admissions probability.

TABLE 10. Probability of Admissions with a One Standard Deviation Score Change Percentage Increase in Brackets

	Average Person	Objective Score Increase	Subjective Score Increase
	(1)	(2)	(3)
Without Correlation	0.207	0.252 (21%)	0.320 (54%)
With Correlation	0.207	0.363 (75%)	0.387 (86%)

These results illustrate that the evaluations provided by the references are more important than the information contained in the objective scores. One has to be cautious before concluding that the objective score plays a minor role. In particular, references may be aware of these scores before they complete their assessment. Second, given the potential for cheap talk, we need to evaluate the reliability of the subjective evaluations, a question we address in Section 7.

7. The Reliability of Subjective Evaluations

The analysis thus far has shown that relative to random selection, the probability that an applicant in our pool is admitted to a top-ten program is highly correlated with both an objective and a subjective measure of performance. Moreover, although these two measures are correlated, admissions committees place more weight upon the subjective measure. The purpose of this section is to compare the evaluations by different references, both to measure their internal consistency and to see whether different references have different impacts upon admissions. Subjective ratings vary sizably within an application, demonstrating that the same applicant can receive very different assessments from different references. Given that these assessments have an impact upon admissions, it is important to understand the source of the variation and whether there is a *reference effect*—namely, that admissions committees place different weights on different references. There are two reasons why this might be the case. The first of these follows from agency theory, which suggests that the quality of the signal may be higher when a reference can relate a person to others. The second, which is difficult to disentangle from the first, is that admissions committees weigh the opinions of certain references more highly than others.

We provide two graphical illustrations of these relationships as follows. Figure 10 illustrates the relationship between a person's objective score quantile and their reference-level subjective score. We also distinguish between references that write only one letter in our sample (red), and more than one letter (blue). Notice that at *all* objective score levels, we observe subjective ratings that go from the best (0.874) to the worst (0.0). Thus, subjective scores are a moderating influence upon the effect of an objective score based on the GRE scores.

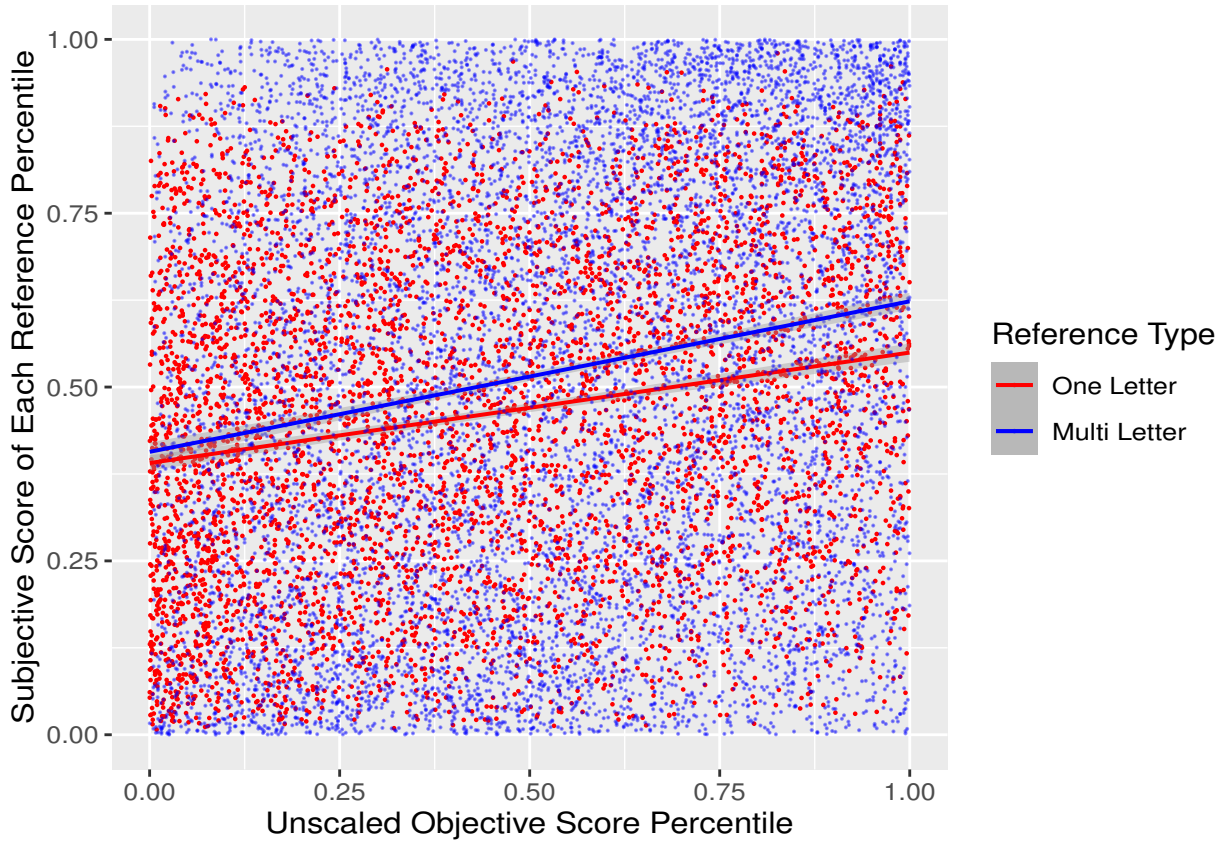


FIGURE 5. Individual Subjective Score Percentile by Objective Score Percentile

We do not know the process that leads a person to choose specific references. Rather than speculate on this process, individuals who choose four or more references are excluded (a small subset of the 6,320 applications). We then code references as submitting one recommendation or 2+ recommendations in our sample. For person i in this sample we let RT_i be the number of references chosen by individual i who appear two or more times in our whole sample. This is a way to capture both reputation effects and the importance of relative performance measures. Second, we plot the linear fit to the data for the two categories of references: one letter and 2+ letters per reference, as shown in table 26. Here we provide a basic control for the endogenous reference choice. In all cases, the subjective score is positively correlated with the objective score. Moreover, the correlation is higher for the references that write more letters, consistent with the hypothesis that these writers are providing more information. However, the fit is not great, consistent with the fact that there is a great deal of variation in assessment. It is possible that the subjective scores are more precise estimates of performance, and one way to evaluate whether they are or not is to examine within-applicant correlation in scores.

Figure 11 plots references' subjective ratings for an applicant against the average subjective rating given by the three references for that applicant. If the references in an application all gave the same scores, then this would be a 45-degree line. Instead, we observe significant heterogeneity for all

TABLE 11. Correlation Between Objective and Individual Level Subjective Score

	Dependent variable: Subjective Score Percentile	
	One Letter	Multi Letter
	(1)	(2)
Objective Score Percentile	0.159	0.216
	(0.011)	(0.011)
Constant	0.391	0.407
	(0.006)	(0.007)
Observations	5,283	9,315
Adjusted R^2	0.040	0.037

Note. Standard errors are heteroskedasticity-robust (HC3).

score levels: references report very different subjective ratings for the same individual. It is this heterogeneity in subjective evaluations that led to the development of standardized tests (Ployhart et al. (2017), Kahneman (2003)). The facts that subjective scores have an impact upon admissions and that the same applicant may receive very different scores from their self-chosen references imply that there is a source of evaluator variation that is out of the control of the applicant. This evaluator variation necessarily leads to noise in the admissions process.

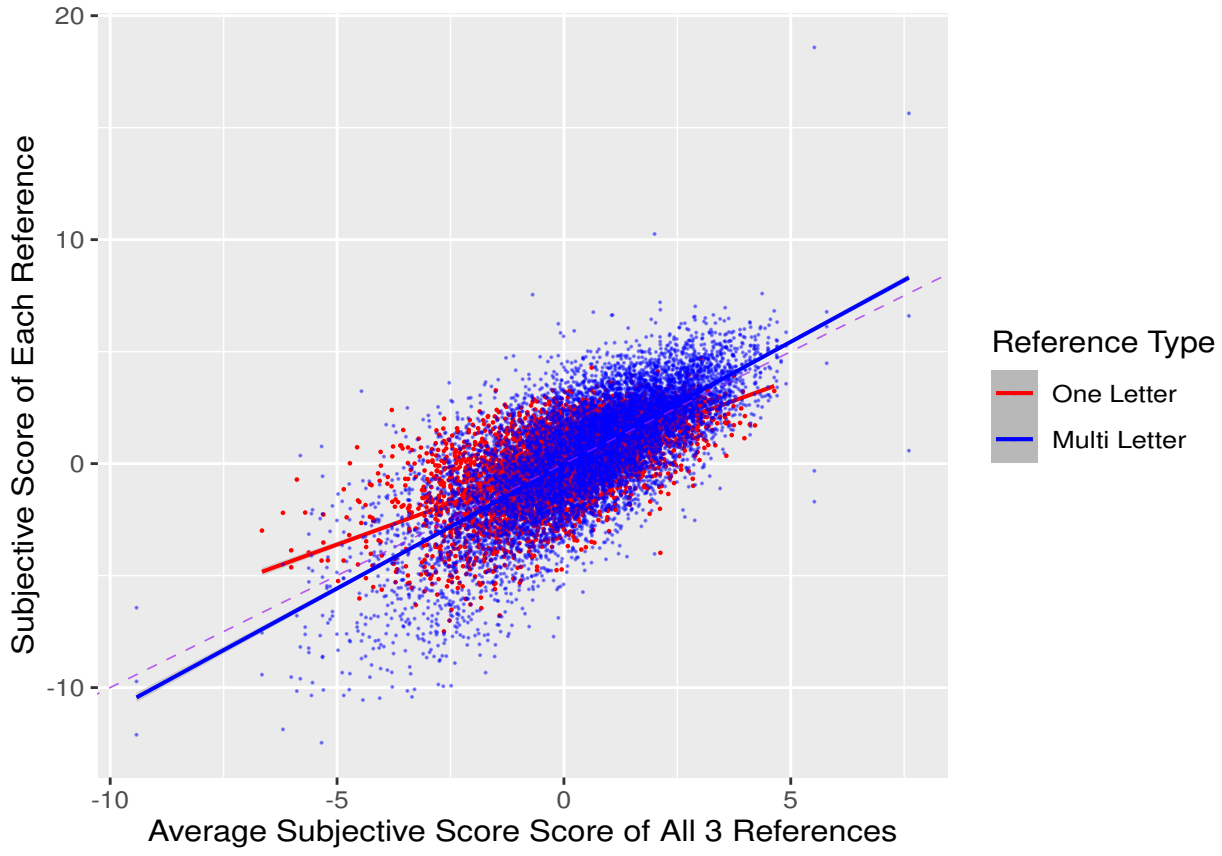


FIGURE 6. Individual Subjective Score versus Applicant Mean Score

The slope of the line of fit in figure 11 is significantly steeper for references who wrote two or more letters (the "2-plus references") than it is for those references who only wrote one. Table 27 reports the results from these regressions. Mechanically, since we regress the value on a mean that includes the variable, the fit is very good. Nonetheless, the difference in slope for single-letter and 2-plus references indicates that references who write more letters discriminate more between applicants. The learning model outlined above implies that if the signal is more informative, then more weight should be placed upon the evaluations from these references. This is something we can test.

TABLE 12. Correlation Between Average Subjective Score and Individual Score

	Dependent variable: Subjective Score Percentile	
	One Letter	Multi Letter
	(1)	(2)
Average Subjective Score	0.734	1.101
	(0.014)	(0.013)
Constant	0.058	−0.068
	(0.019)	(0.022)
Observations	5,283	9,315
Adjusted R^2	0.397	0.542

Note. Standard errors are heteroskedasticity-robust (HC3).

If this hypothesis is correct, then applicants should seek out references from individuals who write more letters. Also, we would likely to see some correlation between the number of letters a reference writes and applicant characteristics. To explore this idea we define RT to be the number of references of an applicant who are prolific, in our case writes 2 or more letters in our sample. Thus, if all the references of an applicant have written only 1 letter, then $RT = 0$. If one reference has written 2 or more letters in our sample, then $RT = 1$ and so on. Since applicants only require three references, then we only consider $RT \in \{0, 1, 2, 3\}$. Table 13 reports application and admission rates by the RT variable and applicants' identity characteristics. All groups, including non-US applicants, have a large number of references who write 2+ letters. This finding is consistent with the observation that many foreign nationals attend universities for undergraduate programs or post-graduate work where they can ask faculty for recommendations. Additionally, admissions probabilities rise with RT_i and Ivy Plus applicants have a larger fraction of 2+ references, consistent with their higher admissions success.

TABLE 13. Identity, Reference Choice and Top 10 Admissions

RT	US Female		US Male		Non-US Female		Non-US Male		Ivy Plus	
	Apply	Success Rate	Apply	Success Rate	Apply	Success Rate	Apply	Success Rate	Apply	Success Rate
0	17%	11%	16%	3%	14%	1%	11%	5%	4%	15%
1	20%	15%	23%	16%	19%	5%	18%	11%	15%	27%
2	31%	33%	30%	25%	31%	16%	32%	20%	33%	39%
3	32%	38%	31%	45%	36%	22%	40%	25%	48%	54%
Total	374	28%	994	26%	1147	14%	2351	19%	461	43%

A natural question is the extent of selection: do individuals with higher RT scores have different performance measures? Table 14 summarizes performance measures grouped by the characteristics of the references and shows strong selection by RT_i. The difference in the mean objective score between $RT = 0$ and $RT = 3$ applicants is 0.81. Since these are z-scored this is 81% of a standard deviation, a very large difference. This result is consistent with less skilled applicants avoiding obtaining references that would place them in the lower part of the reference's distribution.

Notice that the mean objective score of admitted applicants does not depend upon the RT_i variable, yet the subjective score of admitted applicants with $RT_i = 0$ is lower than those with high RT_i scores. This is consistent with the observation that more prolific references are more discriminating, and that candidates a prolific letter writer ranks highly is more likely to be admitted. Admitted students with three 2-plus references have significantly higher subjective scores than admitted students with fewer than three 2-plus references. The last four columns of table (14) report the subjective scores by single letter references and multi-letter references. When $RT_i = 1, 2$ that the applicant has letters from both a prolific and non-prolific reference. Notice that the prolific references systematically rank individuals lower than the individuals who write only one letter, yet the mean score of the admitted applicants is the same. It should be emphasized that the subjective score takes into account all the information we have available, such as the letter features, while the subjective rating (sr) variable is based only upon the responses to the rating questions.

TABLE 14. Performance by Identity, Reference Choice and Top 10 Admissions

	Mean		Mean		Mean Subjective Rating Reference (sd)			
RT	Objective Score (sd)		Subjective Score (sd)		One Letter		Multi Letter	
	Apply	Admitted	Apply	Admitted	Apply	Admitted	Apply	Admitted
0	-0.59 (1.38)	0.54 (0.64)	-0.25 (0.74)	0.38 (0.46)	-0.23 (0.94)	0.37 (0.59)	NaN (NA)	NaN (NA)
1	-0.12 (1.04)	0.51 (0.78)	-0.12 (0.85)	0.51 (0.69)	-0.17 (1.0)	0.35 (0.69)	-0.33 (1.4)	0.35 (0.98)
2	0.1 (0.85)	0.5 (0.69)	-0.04 (1.02)	0.65 (0.72)	-0.14 (1.22)	0.38 (0.94)	-0.23 (1.03)	0.35 (0.73)
3	0.22 (0.78)	0.54 (0.67)	0.19 (1.12)	0.9 (0.86)	NaN (NA)	NaN (NA)	-0.1 (0.86)	0.41 (0.61)
Mean	0.01 (0.99)	0.52 (0.69)	0 (1.01)	0.76 (0.8)	-0.17 (1.1)	0.37 (0.87)	-0.2 (1.06)	0.38 (0.7)

The next question is to ask whether admissions committees place more weight on some references than others. Table 24 can be viewed as an Oaxaca (1973) regression where we ask if objective and subjective scores can control for the self-selection with different reference types. We find strong evidence that the number of 2-plus references has a large and significant correlation with admissions. Thus, while the identity of an applicant seems to have little effect upon admissions after controlling for observed performance, this is not the case with the identity of the reference.

TABLE 15. Reference Effect on Top 10 Admission

	Subsample with Three References				Ivy Plus Subsample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Objective Score - θ^o	0.776	0.748	0.708	0.651	0.907	0.912	0.865
	(0.093)	(0.098)	(0.108)	(0.110)	(0.184)	(0.184)	(0.204)
Subjective Score - θ^S	1.132	1.048	1.037	1.052	1.126	1.073	1.073
	(0.078)	(0.076)	(0.077)	(0.076)	(0.167)	(0.167)	(0.165)
RT = 1		0.759	0.778	0.732		0.688	0.781
		(0.208)	(0.204)	(0.197)		(0.624)	(0.684)
RT = 2		1.366	1.403	1.339		1.102	1.229
		(0.241)	(0.236)	(0.234)		(0.433)	(0.457)
RT = 3		1.527	1.580	1.492		1.536	1.664
		(0.276)	(0.277)	(0.269)		(0.462)	(0.526)
US Female			0.175	0.118			0.054
			(0.136)	(0.136)			(0.250)
Non-US Female			−0.341	−0.180			−0.480
			(0.101)	(0.108)			(0.545)
Non-US Male			−0.156	0.006			−0.544
			(0.146)	(0.162)			(0.322)
Ivy Plus				0.776			
				(0.109)			
Constant	−1.959	−3.159	−3.052	−3.176	−1.335	−2.527	−2.484
	(0.078)	(0.216)	(0.176)	(0.184)	(0.181)	(0.437)	(0.416)
Pseudo R^2	0.2094	0.229	0.2315	0.239	0.219	0.2377	0.2448
Node Purity	0.6748	0.6814	0.6823	0.6848	0.5865	0.594	0.5969
Observations	4,866	4,866	4,866	4,866	461	461	461

Note. Standard errors are clustered by application year (HC3).

It appears that having more references who write multiple letters gives an applicant a step up in admissions probability. The problem with this regression, as we can see from table 14 is that there is positive selection into the set of individuals who obtain recommendations from prolific refernces. We can get some sense of why this may be the case by looking at the reputation of the reference. Table 16 reports the fraction of references that are in the top 10% of the RePEc ranking.[13] In general, the 2-plus references are better known and more highly cited. These results are consistent with the perception that a recommendation letter from a well-known person is likely to have a larger impact.

Notice that graduating from an Ivy Plus college is also positively correlated with admissions, though the point estimate, 0.776, is smaller than a letter from references of type $RT = 2$ or $RT = 1$. Columns (5)-(7) report the results with just the Ivy Plus applicants. The coefficients on both the performance scores and RT variables are similar. This suggests that even within the Ivy Plus group there is significant variation in performance and reference quality.

Applicants with a 2-plus reference have a higher likelihood of acceptance. This is consistent with two reinforcing effects. First, prolific references are more discriminating, and agency theory predicts

[13]From https://ideas.repec.org/top/top.person.all.html, accessed in June 2020. RePEc is a widely used ranking service that documents the productivity of research-active economists.

TABLE 16. Reference Choice and reference Reputation

Number of Letters	RePEc Top 10% Fraction	Objective Score	Subjective Score
1	0.116	-0.216 (1.162)	-0.137 (0.861)
2	0.268	0.054 (0.913)	0.004 (0.968)
3	0.335	0.141 (0.825)	0.057 (0.989)
4	0.480	0.150 (0.850)	0.085 (1.040)
5+	0.540	0.176 (0.794)	0.143 (1.119)

Note. Standard deviation in parentheses.

that a high score from such a reference would carry greater weight. Second, there is evidence of positive selection into prolific references, and hence the quality of candidates with high scores from a prolific reference is likely to be higher than from a less prolific reference. The quality of these recommendations can be further explored by looking at the job placement of graduates.

8. Performance and Job Placement

Assuming that a goal of a PhD program is to produce research faculty, we can use job outcomes to assess the quality of the admissions criteria. In this section, we report results on the relationship between observed applicant performance and academic job placement five to six years later. In particular, doctoral programs tend to view the placement of their graduates into an assistant professor position at either a top-ten or top-twenty department as a success.

Using the data for individuals who are in our data from 2013-2015, we explore the relationship between performance as measured in the application and job placement several years later. Given that the information in the application is not available to the job market, we can assume that conditional upon the program in which the person is accepted, the application data affect outcomes only through the information it contains. These regressions compare the importance of factors that influence admissions with the factors that influence placement with and without conditioning upon program. Table 17 details the placement of students in the 2013-2015 sample and job outcomes. Most top twenty assistant professor jobs in the sample are filled by the graduates of the top-ten departments. Thus, this table illustrates strong sorting between graduates of top departments and those who start their careers in a top department.

TABLE 17. Job Placement Statistics for Applicant Pool 2013-2015

PhD Program Rank	Number Students	Assistant Professor Placement		
		Any	Top 10	Top 20
Top 5	285	119	27	46
Top 6-10	288	112	16	24
Top 11-20	371	112	2	8
Top 21-40	279	74	2	4
Other Econ PhD	382	116	1	2
Other PhDs	851	65	4	10
Total	2456	598	52	94

Table 18 reports the results of a logit regression of observed performance of students in our sample on obtaining an assistant professor position at a top economics department. Columns (1) to (3) report the results for obtaining a top-ten assistant professor position. We can see in column 1 that the objective score is positively correlated with holding a top-ten assistant professorship, and the size of the coefficient is not significantly different from the effect of the objective score on top-ten program admission. This is consistent with the hypothesis that the pathway to a top-ten job is via a top-ten department. Applicants' subjective scores are correlated with top-ten job placement, and the coefficient is significant even after controlling for top-twenty graduate programs. We add indicator variables for attending a top-five, top-ten, and top-twenty departments in the second column. The subjective score retains a significant coefficient, while the objective score is no longer significant. However, attending an Ivy Plus undergraduate college is positively associated with obtaining a top-ten position.

Columns (4)-(6) in table 18 report the results using a top-twenty assistant professor position as an outcome measure. In this case, applicants' objective scores are not significant, while their subjective scores remain significant even after including program dummies. An interesting observation is the negative coefficient for non-US female graduates. This result is consistent with the existence of job opportunities in their home countries that may be preferred to a top-twenty position. However, we cannot exclude the possibility that there is a preference for US citizens for these positions, though there is no evidence of such a preference in the case of top-ten assistant professor positions. Unlike in columns (1)-(2), the coefficients for the higher RT variables are not significant. This is consistent with the hypothesis that prolific references can spot individuals who can obtain top-ten assistant professor positions but their evaluations seem less informative for top-twenty positions.

It is worth discussing the GRE scores' effects since their use has been controversial.[14] In appendix section (A.10), we report the job market success by quantitative GRE ranges. It should be emphasized that when individuals go on the job market, their GRE scores are unknown to the market and so they are not used by hiring committees. Two facts are clear from this table. The first is that only individuals with quantitative GRE scores above the 85th percentile obtained top-ten assistant professor jobs. Second, the number of individuals who obtained top-twenty jobs with scores below the 85th percentile is very small. The numbers in the table are the fraction of individuals admitted in the given GRE range. However, above this cutoff the relationship is non-linear, with individuals in the range of 90-94 having the highest success rate. It is not clear what explains this, but it is consistent with the hypothesis that at the very top these scores are not very informative of research success. Hence, these scores are not predictive of performance conditional upon acceptance into a program. However, the data is not consistent with the hypothesis that the quantitative GRE score has no information, only that is a relatively crude and likely a non-linear measure for the highest-performing individuals. For these individuals, the fact that the subjective scores are correlated with job placement, even after controlling for PhD program, suggests that the subjective evaluations appear to be more useful for the very best individuals.

[14] See American Sociological Association (2021) discussion and recommendation that the GRE exam not be used for sociology admissions.

TABLE 18. Factors Affecting Assistant Professor Placement

	Assistant Professor at:					
	Top 10 Program			Top 20 Program		
	(1)	(2)	(3)	(4)	(5)	(6)
Objective Score - θ^o	0.590	0.511	0.247	0.256	0.196	−0.024
	(0.165)	(0.178)	(0.215)	(0.157)	(0.135)	(0.170)
Subjective Score - θ^S	0.635	0.655	0.342	0.785	0.802	0.510
	(0.012)	(0.010)	(0.044)	(0.048)	(0.056)	(0.132)
RT = 1	−0.573	−0.623	−0.554	0.191	0.147	0.201
	(0.810)	(0.792)	(0.878)	(0.675)	(0.692)	(0.746)
RT = 2	1.297	1.226	0.933	0.811	0.750	0.462
	(0.536)	(0.552)	(0.573)	(0.571)	(0.602)	(0.572)
RT >= 3	1.116	0.988	0.565	0.885	0.770	0.323
	(0.511)	(0.508)	(0.487)	(0.832)	(0.866)	(0.776)
US Female	0.706	0.658	0.560	0.260	0.220	0.126
	(0.563)	(0.565)	(0.608)	(0.545)	(0.542)	(0.550)
Non-US Female	−0.621	−0.455	−0.397	−1.143	−0.993	−0.905
	(0.473)	(0.495)	(0.461)	(0.343)	(0.360)	(0.285)
Non-US Male	−0.378	−0.208	−0.300	−0.540	−0.387	−0.468
	(0.662)	(0.658)	(0.723)	(0.243)	(0.266)	(0.319)
IvyPlus		0.681	0.376		0.667	0.375
		(0.204)	(0.155)		(0.207)	(0.273)
Top 5 Program			1.903			1.961
			(0.231)			(0.535)
Top 6-10 Program			1.571			1.550
			(0.329)			(0.571)
Top 11-20 Program			−0.330			0.280
			(0.655)			(0.772)
Constant	−5.007	−5.104	−5.321	−3.855	−3.957	−4.255
	(0.509)	(0.539)	(0.647)	(0.775)	(0.784)	(0.955)
Pseudo R^2	0.138	0.1442	0.1969	0.1239	0.1299	0.1843
Node Purity	0.9154	0.916	0.9209	0.8674	0.8682	0.8759
Observations	2,456	2,456	2,456	2,456	2,456	2,456

Note. Standard errors are clustered by application year and are heteroskedasticity-robust (HC3).

Overall, these results are consistent with the hypothesis that both objective and subjective performance measures at the time of application to a PhD program are correlated with future success during and after the PhD program as measured by job placement. It is also clear that these effects are mediated by program placement—attending a highly-ranked program is strongly correlated with future job success. Moreover, the identity of the reference seems to matter—the subjective views of individuals who wrote reference letters for more applicants have a greater weight on admissions. If this is pure animus in the Becker sense, then we would expect applicants benefiting from such preferential treatment at the time of admissions to not perform well on the job market. We do not find evidence of this. At the same time, the fact that neither the objective score nor the subjective score has a large effect upon placement, conditional on the quality of the graduate program, highlights the central role of perceived program quality upon job success. Finally, attending an Ivy

Plus undergraduate program is associated with a higher rate of placement into an elite professorial position.

9. Discussion

The admissions process plays a central role in the production of professional economists. It is very dependent upon the good faith participation of human evaluators, both as references for applicants, and as reviewers on admissions committees. The literature on human evaluators points out that data-based models can often perform better than human evaluators (Kahneman (2003); Kahneman and Klein (2009); Dawes and Corrigan (1974)). Our findings may provide some nuance to these results.

9.1. **Main Results.** Despite the difficulties for humans to utilize complex data in the context of admissions, our findings are consistent with the predictions of agency theory, Bayesian learning, and statistical discrimination.[15] All performance measures are found to be correlated with both admissions success and final job placement as an assistant professor at an elite department. These results are conditional upon the pool of individuals who chose to apply to the program that provided data for this study. The LASSO results suggest that the independence assumption for an individual from the application pool is reasonable. Hence, we can suppose that an individual in this pool who obtains higher scores will experience a higher probability of admissions, holding fixed the distribution of applicants.

It is an open question how these estimates would vary should there be a large change in the characteristics of the applicant pool. This is certainly an important question given that graduate programs are experimenting with changes in their admissions requirements. The theory developed in section 3 predicts that if a single program drops the GRE requirement, then they are likely to receive more applicants with lower GRE scores and fewer applicants with higher GRE scores who otherwise have a good chance of admission to higher-ranked programs. However, if individuals choose not to take the GRE, then the theory predicts that their admissions decision is driven by their belief regarding their probability of admissions times their expected benefit versus the cost of applying to a program. Thus, the theory predicts that dropping the GRE requirement will increase applications. It is an open question whether dropping the GRE requirement leads to a better pool of applicants or affects the average quality of admitted students.

Whether or not the GRE score is available, agency theory predicts that relative performance signals provide additional information about applicant quality. However, the same theory also predicts that if the GRE score is not available, then committees will place more weight upon both the level and the relative levels of the subjective ratings. In addition, the fact that Ivy Plus applicants are more successful as a group is consistent with statistical discrimination, as predicted in the sorting model of MacLeod and Urquiola (2015). As a group, Ivy Plus applicants are positively selected relative to the pool of applicants. Given that the identity of the undergraduate college

[15]There is a large literature in psychology showing that humans cannot do a good job of using that data to categorize people into admit or not (Ashby and Maddox (2005, 2011)). This explains in part why there is a great deal of noise in the admission process.

is known, and it is also known to be correlated with observable performance metrics, then agency theory implies that dropping the GRE score requirement will result in more weight being placed upon college identity. This is likely to increase the magnitude of the Ivy Plus effect.

Our results provide some evidence on Spence (1973) signaling model of education. The Spence model predicts that education reveals skill because the cost of acquiring education is lower for high-skill individuals, hence only high-skill individuals acquire additional training. In our case, submitting an application is expensive, implying that merely applying to a program is a positive signal of skill. This is consistent with the fact that there are individuals from all parts of the GRE distribution who are admitted. Individuals with lower GRE scores who apply are signaling that they have characteristics other than a high GRE score that make them suitable for a career in economics.

Another potential signal comes from the reference letter length. Letter length is expensive and effortful, hence one might hypothesize that references signal applicant quality via longer letters. Letter length is positively correlated with admissions rates. However, we cannot conclude that the mechanism arise due to signaling by the reference because the references do not appear to adjust their letter length as a function of applicant quality.

9.2. **Implications for Stratification and Structural Discrimination.** The results also provide some insights regarding career stratification. The results highlight two interlocking effects that are potential mechanisms for explaining structural discrimination. First, the fact that references face a capacity constraint implies that applicants who lack access to a multiple-applicant reference are at a disadvantage. If access to such references is correlated with a person's resources, then we would expect a correlation between those resources and access to the best programs.

The evidence in Clauset et al. (2015) and Stansbury and Schultz (2022) is consistent with these findings. In economics, as Fourcade et al. (2015) observe, there is wide agreement within the profession on how to rank individuals and departments. These rankings get reinforced in a hierarchical sorting system. If elite undergraduate colleges select individuals in a way that is consistent with the rankings of the profession, then Bayesian decision making implies that individuals from these colleges should be favored by admissions committees. The fact that our performance measures are constructed using observed admissions, and the measures are positively correlated with future job market success is consistent with the hypothesis that applicants are chosen based upon an expectation of future performance in the academic market.

Bayesian learning theory predicts that providing more and higher quality signals regarding individual skill increases the admissions chances of individuals who do not have access to prolific references or an Ivy Plus undergraduate degree. For example, the government of Colombia was concerned that students who were not able to attend a flagship university were at a disadvantage. The government addressed this problem by introducing a national college exit exam that high-performing students from lower-ranked programs could use to signal their quality to the labor market. MacLeod et al. (2017) show that this reform led to a reduction in the importance of graduating from an elite program for individuals entering the Colombian labor market.

These results highlight the potential role of information networks and mentorship, especially since jobs in academia occur after a long climb up the education hierarchy. Recent work by Chetty et al. (2022a) and Chetty et al. (2022b) shows that network effects begin early and can occur via social contacts (in their case, friending on Facebook). For individuals from disadvantaged backgrounds, having a connection with more advantaged persons may lead to better labor market outcomes. Such results reinforce an earlier literature that highlights the importance of mentors. Talented students who, for various reasons, are not admitted to an elite undergraduate college would benefit from access to mentors who can provide credible evaluations of their skills and help them in their career choices (Blau et al. (2010), Bayer and Rouse (2016), Porter and Serra (2020) and Ginther et al. (2020)).

Work by Black et al. (2020) and Bleemer (2021) show that policies that increase access to elite undergraduate colleges also lead to improved outcomes for individuals. This raises the interesting question of what the best policy for diversity is in a hierarchical sorting system. If diversity is increased at the elite college level, then preferential treatment of those graduates would be consistent with increasing diversity. Moreover, our findings also provide some evidence on the mechanism at work. Individuals at elite colleges who have access to references who are better known scholars and who write more letters for graduate school are likely to have an advantage in the admissions process. However, since graduate programs are open to applicants worldwide—including to those without access to high quality evaluations—there remains the question of how best to select applicants from this much wider pool.

References

Abel, M., R. Burger, and P. Piraino (2020, July). The value of reference letters: Experimental evidence from south africa. *Am. Econ. J. Appl. Econ. 12*(3), 40–71.

Acemoglu, D. and D. Autor (2011). Chapter 12 - skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 4, Part B, pp. 1043–1171. Elsevier.

American Sociological Association (2021, March). Reconsidering the use of the graduate record examination (gre) in graduate school admissions decisions.

Ashby, F. G. and W. T. Maddox (2005). Human category learning. *Annu. Rev. Psychol. 56*(1), 149–178. PMID: 15709932.

Ashby, F. G. and W. T. Maddox (2011). Human category learning 2.0. *Ann. N. Y. Acad. Sci. 1224*(1), 147–161.

Athey, S., L. F. Katz, A. B. Krueger, S. Levitt, and J. Poterba (2007, April). What does performance in graduate school predict? graduate economics education and student outcomes. *American Economic Review 97*(2), 512–518.

Bayer, A. and C. E. Rouse (2016, November). Diversity in the economics profession: A new attack on an old problem. *Journal of Economic Perspectives 30*(4), 221–242.

Beaman, L., N. Keleher, and J. Magruder (2018). Do job networks disadvantage women? evidence from a recruitment experiment in malawi. *J. Labor Econ. 36*(1), 121–157.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *J. Econ. Perspect. 28*(2), 29–50.

Black, S., J. Denning, and J. Rothstein (2020, March). Winners and losers? the effect of gaining and losing access to selective colleges on education and labor market outcomes. Technical Report w26821, National Bureau of Economic Research, Cambridge, MA.

Blau, F. D., J. M. Currie, R. T. A. Croson, and D. K. Ginther (2010, May). Can mentoring help female assistant professors? interim results from a randomized trial. *Am. Econ. Rev. 100*(2), 348–52.

Bleemer, Z. (2021, January). Top percent policies and the return to postsecondary selectivity.

Brown, M., E. Setren, and G. Topa (2016). Do informal referrals lead to better matches? evidence from a firm's employee referral system. *J. Labor Econ. 34*(1), 161–209.

Chetty, R., M. O. Jackson, T. Kuchler, J. Stroebel, N. Hendren, R. B. Fluegge, S. Gong, F. Gonzalez, A. Grondin, M. Jacob, D. Johnston, M. Koenen, E. Laguna-Muggenburg, F. Mudekereza, T. Rutter, N. Thor, W. Townsend, R. Zhang, M. Bailey, P. Barberá, M. Bhole, and N. Wernerfelt (2022a, August). Social capital i: Measurement and associations with economic mobility. *Nature 608*(7921), 108–121.

Chetty, R., M. O. Jackson, T. Kuchler, J. Stroebel, N. Hendren, R. B. Fluegge, S. Gong, F. Gonzalez, A. Grondin, M. Jacob, D. Johnston, M. Koenen, E. Laguna-Muggenburg, F. Mudekereza, T. Rutter, N. Thor, W. Townsend, R. Zhang, M. Bailey, P. Barberá, M. Bhole, and N. Wernerfelt (2022b, August). Social capital ii: Determinants of economic connectedness. *Nature 608*(7921), 122–134.

Clauset, A., S. Arbesman, and D. B. Larremore (2015, February). Systematic inequality and hierarchy in faculty hiring networks. *Sci. Adv. 1*(1).

Crawford, V. P. and J. Sobel (1982). Strategic information-transmission. *Econometrica 50*(6), 1431–1451.

Dawes, R. and B. Corrigan (1974). Linear-models in decision-making. *Psychol. Bull. 81*(2), 95–106.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *Am. Psychol. 26*(2), 180–188.

DeGroot, M. H. (1972). *Optimal Statistical Decisions*. New York, NY: McGraw-Hill Book C.

Dustmann, C., A. Glitz, U. Schonberg, and H. Brucker (2016). Referral-based job search networks. *Rev. Econ. Stud. 83*(2), 514–546.

Dutt, K., D. Pfaff, A. Bernstein, J. Dillard, and C. Block (2016). Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nat. Geosci. 9*, 805–808.

Estrella, A. (1998). A new measure of fit for equations with dichotomous dependent variables. *J. Bus. Econ. Stat. 16*(2), 198–205.

Farrell, J. and M. Rabin (1996). Cheap talk. *J. Econ. Perspect. 10*(3), 103–118.

Fortin, N., T. Lemieux, and S. Firpo (2010). Decomposition methods. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics, Volume 4*, Volume 4. Elsevier.

Fourcade, M., E. Ollion, and Y. Algan (2015). The superiority of economists. *J. Econ. Perspect. 29*(1), pp. 89–113.

Frederiksen, A., F. Lange, and B. Kriechel (2017). Subjective performance evaluations and employee careers. *J. Econ. Behav. Organ. 134*(C), 408–429.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *J. Econ. Lit. 57*(3), 535–74.

Ginther, D. K., J. M. Currie, F. D. Blau, and R. T. A. Croson (2020, May). Can mentoring help female assistant professors in economics? an evaluation by randomized trial. *AEA Papers and Proceedings 110*, 205–209.

Goldin, C. and C. Rouse (2000, September). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *Am. Econ. Rev. 90*(4), 715–41.

Goodman, J., O. Gurantz, and J. Smith (2020, May). Take two! sat retaking and college enrollment gaps. *Am. Econ. J.-Econ. Policy 12*(2), 115–158.

Goodman, S. (2016). Learning from the test: Raising selective college enrollment by providing information. *Rev. Econ. Stat. 98*(4), 671–684.

Green, J. R. and N. L. Stokey (1983, June). A comparison of tournaments and contracts. *J. Polit. Econ. 91*(3), 349–364.

Grove, W. A. and S. Wu (2007, May). The search for economics talent: Doctoral completion and research productivity. *Am. Econ. Rev. 97*(2), 506–511.

Harris, M. and A. Raviv (1979). Optimal incentive contracts with imperfect information. *J. Econ. Theory 20*, 231–259.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. New York, NY: Springer.

Holmström, B. (1979). Moral hazard and observability. *Bell J. Econ. 10*(1), 74–91.

Hoxby, C. and C. Avery (2013). The missing "one-offs": The hidden supply of high-achieving, low-income students. *Brook. Pap. Econ. Act.*, 1–65. Hoxby, Caroline Avery, Christopher.

Issacc, C., J. Chertoff, B. Lee, and M. Carnes (2011, January). Do students' and authors' genders affect evaluations? alinguistic analysis of medical student performance evaluations. *Acad. Medicne 86*(1), 59–66.

Jacob, B. A. and L. Lefgren (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *J. Labor Econ. 26*(1), 101–136.

Jeganathan, S., S. Parthasarathy, A. R. Lakshminarayanan, P. M. Ashok Kumar, and M. K. A. Khan (2021). Predicting the post graduate admissions using classification techniques. In *2021 Int. Conf. Emerg. Smart Comput. Inform. ESCI*, pp. 346–350.

Kahneman, D. (2003). Les prix nobel 2002. Chapter Autobiography. Stockholm, Sweden: Almqvist & Wiksell International. This has the citation to Kahneman settuign up system to interview pilots and the results.

Kahneman, D. and G. Klein (2009). Conditions for intuitive expertise: A failure to disagree. *Am. Psychol. 64*(6), 515–526.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions*. *Q. J. Econ. 133*(1), 237–293.

Krueger, A. and S. Wu (2000). Forecasting job placements of economics graduate students. *J. Econ. Educ. 31*(1), 81–94.

Kuncel, N. R., S. Wee, L. Serafin, and S. A. Hezlett (2010). The validity of the graduate record examination for master's and doctoral programs: A meta-analytic investigation. *Educ. Psychol. Meas. 70*(2), 340–352.

Lester, B. R., D. Rivers, and G. Topa (2021, October). The heterogeneous impact of referrals on labor market outcomes. Technical Report 987, Federal Reserve Bank of New York, New York, NY.

Lu, V. E. and D. T. Tsotsong (2021, January). Harvard college receives record-high 57,000 applications, delays admissions release date | news | the harvard crimson. *THe Harvard Crimson*.

Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. John Wiley & Sons.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In X. Chen and N. R. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr*, pp. 437–461. New York, NY: Springer.

MacLeod, W. B., E. Riehl, J. E. Saavedra, and M. Urquiola (2017, July). The big sort: College reputation and labor market outcomes. *Am. Econ. J. Appl. Econ. 9*(3), 223–61.

MacLeod, W. B. and M. Urquiola (2015). Reputation and school competition. *Am. Econ. Rev. 105*(11), 3471–3488.

Madera, J., M. Hebl, and R. Martin (2009, November). Gender and letters of recommendation for academia: Agentic and communal differences. *J. Appl. Psychol. 94*, 1591–9.

Milkovich, G. T., J. M. Newman, and C. Milkovich (2017). *Compensation* (Twelfth ed.). McGraw-Hill.

Mulhern, C. (2020). Changing college choices with personalized admissions information at scale: Evidence on naviance. *J. Labor Econ.*.

Murnane, R. (1975). *The Impact of School Resources on the Learning of Inner-City Children*. Ballinger.

Oaxaca, R. (1973). Male-female wages differentials in urban labor markets. *Int. Econ. Rev. 14*, 693–709.

Pallais, A. and E. G. Sands (2016, December). Why the referential treatment? evidence from field experiments on referrals. *J. Polit. Econ. 124*(6), 1793–1828.

Ployhart, R. E., N. Schmitt, and N. T. Tippins (2017). Solving the supreme problem: 100 years of selection and recruitment at the journal of applied psychology. *J. Appl. Psychol. 102*(3), 291–304.

Porter, C. and D. Serra (2020, July). Gender differences in the choice of major: The importance of female role models. *Am. Econ. J. Appl. Econ. 12*(3), 226–254.

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika 20*(2), 93–111.

Rokach, L. and O. Maimon (2005, November). Top-down induction of decision trees. *IEEE Trans. Syst. Man Cybern. 35*(4), 476–487. discusses measures of node impurity and has a general impurity function.

Spence, M. (1973, August). Job market signaling. *Q. J. Econ. 87*(3), 355–374.

Stansbury, A. and R. Schultz (2022). Socioeconomic diversity of economics phds. *SSRN Journal*.

Stock, W. A. and J. J. Siegfried (2014, October). Fifteen years of research on graduate education in economics: What have we learned? *J. Econ. Educ. 45*(4), 287–303.

Topel, R. H. and M. P. Ward (1992). Job mobility and the careers of young men. *Q. J. Econ. 107*(2), 439–479.

Trix, F. and C. Psenka (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse Soc. 14*(2), 191–220.

Waters, A. and R. Miikkulainen (SPR 2014). Grade: Machine-learning support for graduate admissions. *AI Mag. 35*(1), 64–75.

Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgement to Calculation*. W. H. Freemand and Company.

Young, L. and S. Soroka (2012, April). Affective news: The automated coding of sentiment in political texts. *Polit. Commun. 29*(2), 205–231.

Appendix A. Online Appendix

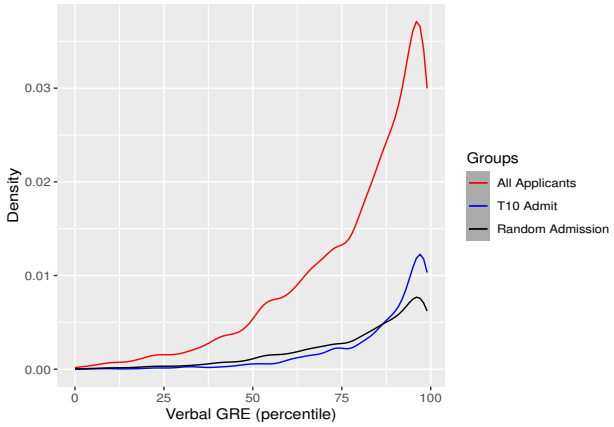


Figure 7. Distribution of GRE Verbal GRE Scores in the Sample

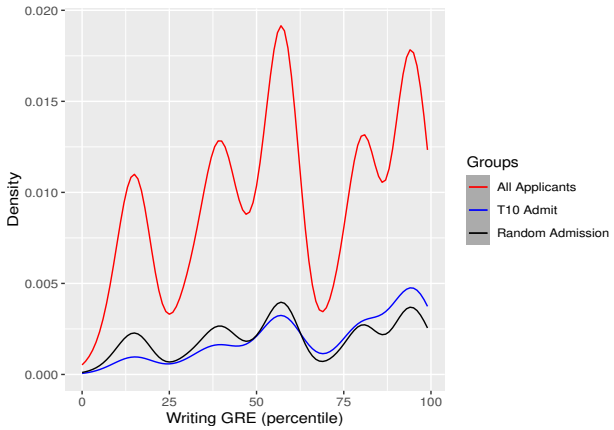


Figure 8. Distribution of GRE Writing GRE Scores in the Sample

A.1. **GRE Scores and Variable Construction for Non-Linearities.**

A.2. **Proof of Proposition (1).** The optimal policy can be formulated as finding a measurable function $\delta : \Re^k \to \Re$ to solve:

$$max_{\delta \in \Delta} W(\delta) = \int_{\omega \in \Re^k} \delta(\omega) p(\hat{\alpha}(\omega)) f(\omega) d\omega, \tag{14}$$

subject to:

$$A(\delta) = \int_{\omega \in \Re^k} \delta(\omega) f(\omega) d\omega \leq r_a, \tag{15}$$

$$\delta(\omega) \leq 1, \tag{16}$$

$$\delta(\omega) \geq 0. \tag{17}$$

Ignoring for the moment, constraints (16-17), the Lagrangian for the optimal solution is given by:

$$\begin{aligned} L(\delta, \lambda) &= W(\delta) + \lambda (r_a - A(\delta)), \\ &= \int_{\omega \in \Re^k} \delta(\omega) (p(\hat{\alpha}(\omega)) - \lambda) f(\omega) d\omega + \lambda r_a. \end{aligned}$$

This is an infinite-dimensional optimization problem whose solution will satisfy the necessary conditions for an optimum (see Luenberger (1969)). Maximizing this subject to constraints (16-17) implies that the optimal rule must satisfy (3). From this we can define two functions of λ:

(1) The upper bound function:

$$r^H(\lambda) = \int_{\{\omega \in \Re^k | p(\hat{\alpha}(\omega)) \geq \lambda\}} f(\omega) d\omega \in [0, 1],$$

is the fraction of individuals whose probability of success is greater than or equal to λ. Since $p(\cdot)$ is increasing in skill, this function is decreasing in λ as the number of persons who satisfy the criteria must be decreasing.

(2) The lower bound function:

$$r^L(\lambda) = 1 - \int_{\{\omega \in \Re^k | p(\hat{\alpha}(\omega)) \leq \lambda\}} f(\omega) d\omega \in [0, 1],$$

is the fraction of individuals whose probability of success is strictly greater than λ (or less than or equal to λ).

Notice that both functions are decreasing in λ with $1 = r^H(0) \geq r^H(\lambda) \geq r^L(\lambda) \geq r^L(1) = 0$. Whenever the probability of success is exactly $\lambda = 0$, then $r^H(\lambda) = r^L(\lambda)$. When this does not occur, there can be a group of students that has the same probability of admissions. In such a case, one might have to randomize admissions over this group.

Next, let $\lambda^H = sup\{\lambda | r^H(\lambda) \geq r^a\}$. By the Lebesgue dominated convergence theorem, $r(\lambda^H) \geq r^a$. Similarly, let $\lambda^L = inf\{\lambda | r^L(\lambda) \leq r^a\}$. Thus, we have:

$$r^H(\lambda^H) \geq r^a \geq r^L(\lambda^L),$$

and it follows that $\lambda^H = \lambda^L$. The optimal $\lambda^* = \lambda^H = \lambda^L$. If the set of applicants on the margins of admissions is of measure zero, $\int_{\{\omega \in \Re^k | p(\hat{\alpha}(\omega)) = \lambda^*\}}$ then $r^H(\lambda^*) = r^a = r^L(\lambda^*)$ and there exists an almost everywhere unique solution to (3). Otherwise, any solution that satisfies (3) and uses some method to choose a subset of students satisfying $p(\hat{\alpha}(\omega)) = \lambda^*$ is optimal.

A.3. **Proof of Proposition 1.** Under the assumption of full support for $f(\cdot)$ it is straightforward to show that $A(\delta^*(\cdot|\lambda))$ is increasing in λ, and hence, if it is the optimal solution, it must be the case that:

$$r_a - \int_{\omega \in \Re^k} \delta(\omega) f(\omega) d\omega = 0. \tag{18}$$

If not, one could admit more students and increase the payoff. Let λ^* be the value that achieves this. A similar argument implies that any other potentially optimal solution, say δ', must also satisfy (18), otherwise it could add more students and increase the payoff. Let $\Omega^A \in \Re^2$ be the set of admitted students and Ω^R the set of rejected students under δ^*. Suppose that δ^* is not optimal and that δ' has a higher payoff. Then it must reject a non-zero number of applicants in the set Ω^A. Given that both policies choose the same number of students, we have:

$$\int_{\{g,s\} \in \Omega^A} \left\{\delta(g,s) - \delta'(g,s)\right\} f(g,s) dsdg = \int_{\{g,s\} \in \Omega^R} \delta'(g,s) f(g,s) dsdg > 0.$$

Hence, it must be the case that:

$$\begin{aligned}
\int_{\{g,s\} \in \Omega^A} \left\{\delta(g,s) - \delta'(g,s)\right\} y(\hat{\alpha}(g,s)) f(g,s) dsdg &\geq \lambda^* \int_{\{g,s\} \in \Omega^A} \left\{\delta(g,s) - \delta'(g,s)\right\} f(g,s) dsdg, \\
&= \lambda^* \int_{\{g,s\} \in \Omega^C} \delta'(g,s) f(g,s) dsdg, \\
&> \int_{\{g,s\} \in \Omega^R} \delta'(g,s) y(\hat{\alpha}(g,s)) f(g,s) dsdg.
\end{aligned}$$

The latter inequality follows from the fact that $y < \lambda^*$ in the set Ω^R. Thus, δ' cannot be optimal, and $\delta^*(\cdot|\lambda^*)$ is an optimal admissions policy.

A.4. **Node Purity.** The node purity for individual i is defined as:

$$\begin{aligned}
NP_i &= E\{a_i = \delta_i | x_i\}, \\
&= a_i \times p_i + (1 - a_i) \times (1 - p_i).
\end{aligned}$$

When $NP_i = 1$ the modelperfectly predicts outcomes, and when $NP_i = 0$ then the model is perfectly wrong. In the machine learning literature, the measure $1 - NP_i$ is called node impurity. This measure is used as a criterion for constructing decision trees (see Rokach and Maimon (2005) for a review). Conveniently, the node purity is directly related to the value of the likelihood function.

The node purity of the model for individuals $i \in I$ and estimated coefficients $\hat{\beta}$ is given by:

$$\begin{aligned} P\left(I, \hat{\beta}\right) &\equiv L\left(I, \hat{\beta}\right)^{1/n} \\ &= \left(\prod_{i \in I} a_i p_i + (1 - a_i)(1 - p_i)\right)^{1/n}, \\ &= gm\left(\vec{NP}\right) \end{aligned}$$

where n is the number of individuals in I, $L(\cdot)$ is the likelihood function, $\vec{NP} = \{NP_i\}_{i \in I}$ is the vector of individual node purities, and the function $gm()$ is the *geometric mean* defined by:

$$\mu = gm(\vec{p}) = \left(\prod_{i \in I^0} p_i\right)^{1/n^0} \in [0, 1]. \tag{19}$$

Thus, node purity is an increasing function of the likelihood function, and the maximum likelihood estimate is also the model that maximizes node purity, as defined in the machine learning literature. Hence, node purity can be mapped one-to-one to the value of the likelihood function, with the difference that its value is very easy to interpret since it lies in the interval $[0, 1]$. In our tables, we report the node purity rather than the value of the likelihood.

Observe that $P\left(I, \hat{\beta}\right) = 1$ if and only if the model predicts perfectly, and hence 1 is the upper bound to the model in all cases. Suppose that one has no information regarding an individual's characteristics. In that case, the probability that a person is admitted is given by the number of admits divided by the number of applicants. Let

$$\underline{p}(I) = \frac{\#\{i \in I | a_i = 1\}}{\#\{i \in I\}},$$

be the unconditional probability of admissions. This is the maximum likelihood estimate when there are no covariates and therefore provides a *lower bound* on node purity. This observation motivates Estrella (1998) to define a pseudo R^2 measure:

$$\begin{aligned} \phi\left(I, \hat{\beta}\right) &= \left(1 - \left(\frac{log\left(\underline{p}(I)\right)}{log\left(P\left(I, \hat{\beta}\right)\right)}\right)^{2log(\underline{p}(I))}\right). \\ &= \left(1 - \lambda^{2log(\underline{p}(I))}\right), \end{aligned}$$

where $\lambda = log\left(\underline{p}(I)\right) / log\left(P\left(I, \hat{\beta}\right)\right) \geq 1$ is the likelihood ratio between the constrained and unconstrained models. Under the null hypothesis that $\hat{\beta}$ are zero, this yields the well-known likelihood ratio test from Rao (1955):

$$t \equiv \frac{\left(1 - \phi\left(I, \hat{\beta}\right)\right)}{log\left(\underline{p}(I)\right)} \sim \chi^2(k),$$

where k is the number of variables in $\hat{\beta}$. Thus, this goodness of fit lies on the unit interval and can be used to construct a valid test statistic to assess model fit.

A.5. **Outcome Variable Definitions and Top Economics Programs.** We define top five, ten, and twenty based on the US News ranking of economics programs.[16] For the top ten admission outcome variable, an application is coded as T10 if we observe the applicant attend a top-ten school for a PhD with a program title that includes either econ* or business and is the last application from that applicant in our data. The * indicates that programs containing the letter string "econ" are included. For the assistant professor outcome variables, an individual is coded as an assistant professor if we observe that they hold a post-PhD job with the title "Assistant Professor." Top ten department includes Top 5 and Top 6-10; top twenty includes Top 5, Top 6-10, and Top 11-20, a ranking that is remarkably stable over time.

Top 5: Harvard University, Massachusetts Institute of Technology, Princeton University, Stanford University, University of California Berkeley

Top 6-10: Columbia University, Northwestern University, University of Chicago, University of Pennsylvania, Yale University

Top 11-20: Brown University, Cornell University, Duke University, London School of Economics, New York University, University of California Los Angeles, University of California San Diego, University of Michigan, University of Minnesota, University of Wisconsin

Top 21-40: Boston College, Boston University, Carnegie Mellon University, Johns Hopkins University, Michigan State University, Ohio State University, Pennsylvania State University, Texas A&M University, University of California Davis, University of California Santa Barbara, University of Illinois Urbana-Champaign, University of Maryland, University of North Carolina, University of Rochester, University of Southern California, University of Texas at Austin, University of Virginia, University of Washington, Vanderbilt University, Washington University in St. Louis

Note that the definition of the top ten admission outcome variable includes any program, such as the business school, at a top 10 university. However, looking only at directories from top-ten programs, 69.6% of students from 2013 to 2019 are in our sample, and hence about 30% of the admitted students to a top 10 program did not apply to the program used in this study. In the three cases where the student directory does not include first year of attendance, we approximate by using the directories from the first half of both 2019 and 2020. This assumes a six-year PhD completion time. Excluding the program that provided the data, 66.0% of students in the directories of top-ten programs are in our sample. When grouping by program, the percent in our sample for the lowest, median, and highest program is 55.7%, 66.1%, and 74.8%, respectively.

A.6. **Dictionaries.** This section lists the full dictionaries used for the topic analysis used to construct the subjective score used in the next section. A * indicates that words containing the letter string that precedes the asterisk are included.

Standout words:

excellen*, superb, outstanding, unique, strongest, exceptional, unparalleled, unusual, genius, brillian*, perfect, wonderful, terrific*, fabulous, magnificent, remarkable, extraordinar*, amazing, supreme*, unmatched, unprecedented

Ability words:

[16] See US News Best Economics Schools.

talent*, intell*, smart*, skill*, ability, bright*, brain*, aptitude, gift*, capacity, propensity, innate, flair, knack, clever*, expert*, proficient*, capable, adept*, able, competent, natural*, inherent*, instinct*, adroit*, creative*, insight*, analytical, intuiti*

Grindstone words:

depend*, diligen*, reliab*, effort*, trust*, responsib*, work*, persist*, organiz*, hardworking, conscientious, meticulous, passion, thorough, dedicate*, careful*, assiduous, methodical, industrious, busy, disciplined, enthusiastic, drive*, "determin*, motivat*

Research words:

research*, data, studies, experiment*, scholarship, test*, code, result*, finding*, model*, publication*, publish*, vita*, method*, scien*, grant*, fund*, manuscript*, project*, original, productive, journal*, theor*, discover*, contribution*, conduct*

Communal words:

affectionate*, helpful, kind*, sympathetic*, sensitiv*, nurtur*, agreeable, tact*, interpersonal, warm*, sweet, caring, gentle, friend*, personable, nice, pleasant, congenial, compassion*, personality, characteristics, qualities, cordial, outgoing, sociable, gregarious, amicable, good

Positive words: From the Lexicoder Sentiment Dictionary, consisting of 1,709 "positive" sentiment words.

Negative words: From the Lexicoder Sentiment Dictionary, consisting of 2,858 "negative" sentiment words.

A.7. **LASSO Appendix.** In this section, we show that our multivariate logit model has external validity in predicting admissions. We use a five-fold LASSO regression to show that the machine-learning-selected best model performs similarly to the simple multivariate logit model.

A.7.1. *Literature.* The Least Absolute Shrinkage and Selection Operator (LASSO) performs variable selection under high dimensionality and has become increasingly popular in econometrics (Belloni et al. (2014)). Many empirical studies use the LASSO method to test the robustness of the main results. In a context similar to our paper, Abel et al. (2020) use the double-selection post-LASSO method to study the effect of reference letters on employment in South Africa. **?** use a five-fold cross-validated LASSO logistic model to identify words from EJMR posts that are most indicative of the male and female gender. The wide usage of LASSO as a robustness test method in labor economics motivates us to do the same. We examine whether our main results still hold when only using LASSO-selected variables as predictors of admissions.

A.7.2. *Procedure.* First, we use the LASSO model and five-fold random cross-validation to select variables that are predictive of the top ten admission outcomes for the whole applicant pool. To implement the LASSO regression, we randomly split the dataset into five equally-sized and mutually exclusive folds. We then create a LASSO model by selecting one fold as the testing dataset while the remaining four folds serve as the training dataset. We compute the within-sample and out-of-sample goodness of fit of this training model. We use node purity as defined in this paper as a goodness-of-fit measure that builds on log-likelihood. This process is repeated for each of the five

folds. We also estimate the standard errors of node purity by computing the standard deviations of the five folds.

As discussed in the text, the independent variables for the LASSO training are: GRE test scores and quantiles, subjective rating variables, letter length, and letter page count indicator variables, and letters text features, which are word counts normalized by the letter length. All of the independent variables are scaled to mean zero and standard deviation one.

The dependent variable is an indicator variable of whether the applicant gains admission to a top-ten economics PhD program.

TABLE 19. 5-Fold LASSO Top 10 University In- & Out-of-sample Node Purity Table

alpha	lambda	out-of-sample mean	out-of-sample SD	in-sample mean	in-sample SD
1	0.1000	0.6091	0.0162	0.6095	0.0031
1	0.0750	0.6203	0.0154	0.6209	0.0032
1	0.0562	0.6292	0.0151	0.6304	0.0030
1	0.0422	0.6386	0.0141	0.6401	0.0030
1	0.0316	0.6450	0.0134	0.6466	0.0031
1	0.0237	0.6499	0.0129	0.6521	0.0032
1	0.0178	0.6542	0.0130	0.6570	0.0032
1	0.0133	0.6574	0.0130	0.6604	0.0032
1	0.0100	0.6594	0.0129	0.6629	0.0033
1	0.0075	0.6608	0.0130	0.6648	0.0033
1	0.0056	0.6617	0.0131	0.6663	0.0032
1	0.0042	0.6622	0.0131	0.6673	0.0032
1	0.0032	0.6625	0.0131	0.6680	0.0032
1	0.0024	0.6626	0.0131	0.6685	0.0032
1	0.0018	0.6627	0.0131	0.6689	0.0032
1	0.0013	0.6626	0.0131	0.6691	0.0032
1	0.0010	0.6625	0.0131	0.6692	0.0032
1	0.0007	0.6624	0.0131	0.6693	0.0032
1	0.0006	0.6623	0.0131	0.6694	0.0032
1	0.0004	0.6623	0.0132	0.6695	0.0032
1	0.0003	0.6623	0.0132	0.6696	0.0032
1	0.0002	0.6622	0.0132	0.6696	0.0032
1	0.0002	0.6622	0.0132	0.6696	0.0032
1	0.0001	0.6621	0.0132	0.6696	0.0032
1	0.0001	0.6621	0.0133	0.6697	0.0032
1	0.0001	0.6620	0.0132	0.6697	0.0032
1	0.0001	0.6620	0.0132	0.6697	0.0032
1	0.00004	0.6620	0.0132	0.6697	0.0032
1	0.00003	0.6619	0.0132	0.6697	0.0032
1	0.00002	0.6619	0.0132	0.6697	0.0032
1	0.00002	0.6619	0.0132	0.6697	0.0032
1	0.00001	0.6618	0.0132	0.6697	0.0032
1	0.00001	0.6618	0.0132	0.6697	0.0032
1	0	0.6618	0.0132	0.6697	0.0032

A.7.3. *Identification Discussion.*

Observation 1: The out-of-sample fit of the multivariable logit model and that of the best lambda LASSO model are similar.

The alpha-lambda table reports the mean and the standard deviation of the LASSO model's node purity under different tuning parameters λ. Using the node purity measure for goodness of fit, the highest out-of-sample node purity 0.6627 occurs when $\lambda = 0.0018$, which is very close to the node purity 0.6618 at $\lambda = 0$. We perform a Welch two-sample t-test between the two node purity values. The t-statistic is 0.108, and the p-value is 0.92. Therefore, we fail to reject that the out-of-sample fit of the multivariable logit model and that of the best lambda LASSO model are different.

Observation 2: The out-of-sample fit and the within-sample fit of the best lambda LASSO model are quantitatively the same.

We perform a similar Welch t-test between the in- and out-of-sample node purity for the best lambda, $\lambda = 0.0018$, and the t-statistic is 1.02, with a p-value of 0.36. So, the best LASSO model performs similarly both in- and out-of-sample, allowing us to potentially make causal arguments using this model. We use the best lambda model at $\lambda = 0.0018$ for score construction and regression analysis.

Additionally, we perform another Welch t-test between the in- and out-of-sample fit for the $\lambda = 0$ model (our baseline model with all the variables), and the t-statistic is 0.51, with a p-value of 0.63. This shows that a simple multivariate logit model performs similarly in- and out-of-sample. Since $\lambda = 0$ implies a LASSO model without any regularization, this is evidence that we can simply use the multivariable logit regression for prediction and inference.

hese two observations give rise to a strong inference that our best lambda LASSO model is externally valid.

A.7.4. *LASSO Trained Coefficients.* The tables below display coefficients of a set of explanatory variables for top ten graduate economics admissions from our logit regressions. The dependent variable is whether the student was admitted to a top-ten program. The goal of running these regressions is to compute the optimal weights for aggregating the large set of potential explanatory variables into a few variables that explain admissions. This is a dimension-reduction problem, and we want to reduce the "kitchen sink" to a few aggregated scores, which are shown in Table 4. In order to solve the dimension-reduction problem, we use the LASSO method. Tables 20, 21, and 22 report the coefficients under the model (12). Column 1 in Tables 20, 21, and 22 reports the coefficients where all the variables are normalized to a mean of 0 and standard deviation of 1, while column 2 in all of these tables uses the same normalized variables and reports the LASSO coefficients under the cross-validated best lambda model. Therefore, the coefficients in column 2 are biased since our optimal fitted lambda from cross-validation is 0.0018, a value that is different from 0. Also, we choose not to report the standard deviation in column 2 because the goal of LASSO is to select variables, not to evaluate their significance. Furthermore, The current machine learning literature has not yet concluded how to optimally compute LASSO standard deviation. In column 3 of Tables 20, 21 , and 22, we use only the set of variables selected by LASSO in column 2 and run a logit regression without normalizing the selected variables. We do not normalize the variables here because when we create the aggregate scores later, we will normalize these scores to mean 0

and standard deviation 1. If we normalize the variables in column 3 and then normalize them again when creating the aggregate scores, this double normalization could introduce noise to the score aggregation procedure.

TABLE 20. Regression of Top 10 Econ PhD Admission - GRE Results

Panel A: T10 Admission (Part 1 Objective Score Variables)	All Variables (1)	LASSO (2)	Score Weights (3)
GRE Quant Score	0.419 (0.190)	0.534	5.001 (2.236)
GRE Quant Score = 2nd Quartile	0.228 (0.087)	0.109	0.500 (0.195)
GRE Quant Score = 3rd Quartile	0.150 (0.095)	0.033	0.364 (0.245)
GRE Quant Score = Top (4th) Quartile	0.369 (0.128)	0.225	0.789 (0.279)
GRE Verbal Score	0.026 (0.130)	0.092	0.375 (0.416)
GRE Verbal Score = 2nd Quartile	0.036 (0.084)		
GRE Verbal Score = 3rd Quartile	0.184 (0.112)	0.109	0.327 (0.125)
GRE Verbal Score = Top (4th) Quartile	0.218 (0.134)	0.14	0.372 (0.144)
GRE Writing Score	0.368 (0.124)	0.188	0.715 (0.168)
GRE Writing Score = 2nd Quartile	−0.062 (0.066)		
GRE Writing Score = 3rd Quartile	−0.173 (0.106)	-0.014	−0.070 (0.078)
GRE Writing Score = Top (4th) Quartile	−0.202 (0.139)		

Note. Standard errors are heteroskedastic robust.heteroskedasticity-robust (HC3 - pseudo-jackknife).

TABLE 21. Regression of Top 10 Econ PhD Admission - Subjective Evaluation Variables)

	All Variables (1)	LASSO Estimates (2)	Score Weights (3)
Academic Performance = 1	0.234 (0.411)	0.316	0.936 (0.172)
Academic Performance = 2	−0.064 (0.336)		
Academic Performance = 3	−0.199 (0.247)	-0.121	−0.684 (0.318)
Academic Performance = 4	−0.071 (0.166)	-0.016	−0.211 (0.677)
Academic Performance = 5	−0.148 (0.128)	-0.042	−1.505 (1.490)
Academic Performance = Unknown	−0.071 (0.131)	-0.034	−0.442 (0.448)
Intellectual Potential = 1	−0.246 (0.450)		
Intellectual Potential = 2	−0.206 (0.374)		
Intellectual Potential = 3	−0.218 (0.266)	-0.044	−0.306 (0.357)
Intellectual Potential = 4	−0.256 (0.171)	-0.175	−1.935 (0.920)
Intellectual Potential = 5	0.072 (0.124)		
Intellectual Potential = Unknown	−0.064 (0.089)	-0.013	−0.512 (1.012)
Research Potential = 1	0.603 (0.399)	0.047	−0.017 (0.238)
Research Potential = 2	0.326 (0.325)	-0.126	−0.649 (0.220)
Research Potential = 3	0.369 (0.223)		
Research Potential = 4	0.126 (0.151)	-0.024	−0.461 (0.901)
Research Potential = 5	0.019 (0.123)		
Research Potential = Unknown	0.156 (0.127)	-0.003	−0.233 (0.557)
Writing Skill = 1	−0.174 (0.242)	0.064	0.224 (0.153)
Writing Skill = 2	−0.230 (0.226)		
Writing Skill = 3	−0.288 (0.173)	-0.103	−0.554 (0.237)
Writing Skill = 4	−0.346 (0.118)	-0.213	−1.767 (0.445)
Writing Skill = 5	−0.199 (0.090)	-0.11	−1.956 (0.941)
Writing Skill = Unknown	−0.123 (0.126)		

Note. Standard errors are heteroskedasticity-robust (HC3 - pseudo-jackknife).

TABLE 22. Regression of Top 10 Econ PhD Admission - Letter Features

	All Variables (1)	LASSO Estimates (2)	Score Weights (3)
Letter Length	0.299 (0.067)	0.303	0.002 (0.001)
page count	0.043 (0.086)	0.047	0.098 (0.152)
page count (1, 2]	0.187 (0.089)	0.092	0.408 (0.168)
page count (2, 3]	0.120 (0.096)	0.039	0.307 (0.210)
page count (3, 4]	0.014 (0.035)		
page count 4+	−0.071 (0.077)	-0.069	−1.831 (1.749)
Ability Words	−0.187 (0.042)	-0.169	−0.368 (0.082)
Standout Words	0.227 (0.039)	0.203	0.642 (0.112)
Research Words	0.079 (0.043)	0.056	0.061 (0.032)
Grindstone Words	−0.066 (0.045)	-0.042	−0.096 (0.066)
Teaching Words	−0.008 (0.045)		
Communal Words	−0.109 (0.044)	-0.097	−0.327 (0.135)
Agentic Words	−0.041 (0.038)	-0.02	−0.232 (0.217)
Negative Words	0.072 (0.036)	0.054	0.100 (0.050)
Positive Words	−0.142 (0.047)	-1.133	−0.083 (0.027)
Constant	−1.919 (0.055)	-1.853	−8.298 (1.970)
Node Purity	0.6689		0.6686
Observations	6,320	6,320	6,320

Note. All variables in columns 1 and 2 are scaled to mean 0 and standard deviation of 1. Column 1 is a logistic regression of T10 admission outcome on the scaled variables. Column 2 uses LASSO at the optimal lambda value, which was chosen by 5-fold cross-validation method discussed in the appendix. Column 3 runs a logistic regression of T10 admission outcome on the LASSO-selected variables without scaling them.The variable "Academic Performance" corresponds to "Academic performance" in the subjective evaluation form. The variable "Intellectual Potential" corresponds to "Intellectual potential" in the subjective evaluation form. The variable "Research Potential" corresponds to "Graduate research potential" in the subjective evaluation form. The variable "Writing Skill" corresponds to "Command of writing" in the subjective evaluation form. The evaluation scale "1", "2", "3", "4", "5", and "Unknown" correspond to "Exceptional", "Top 5%", "Top 10%", "Top 20%", "Top 50%", and "Unable to Judge" in the subjective evaluation form. Some references left blank, which are coded as "NA" and used as the baseline for indicator variable comparison here. All the word count variables (e.g. ability words) are number of word count in this category normalized by the letter length. Standard errors are heteroskedasticity-robust (HC3 - pseudo-jackknife).

A.8. **Identity and Admissions.** In our sample we have a large number of foreign students as well as a significant number of female applicants. Recall that our scores aggregate the evaluations based only upon performance data and not contain information regarding the identity of the students. We begin by showing that the four populations differ significantly in their mean observable characteristics. We then do a Oaxaca (1973) type decomposition and ask whether controlling for observed characteristics can explain the variation in admissions by group. Table 1 shows that non-US female applicants as a group have the lowest likelihood of admission, while US female applicants have the highest. This illustrates the importance of sub-group variation—asking how only gender affects admissions in this case does not take into account the large disparity in admissions between US and non-US female applicants.

Table 23 shows that both objective and subjective performance measures are lower for non-US female applicants. This is still the case for the subjective scores of admitted students. It is particularly interesting to note that the objective score for the non-US students is lower than for the US students, suggesting that the selection effects are quite different for each group.

TABLE 23. Objective and Subjective Score Summary by Identity Group

	All		Admitted	
	Objective Score	Subjective Score	Objective Score	Subjective Score
US Male	0.291 (0.03)	0.157 (0.025)	0.832 (0.032)	0.727 (0.042)
Non-US Male	-0.036 (0.016)	-0.004 (0.018)	0.365 (0.026)	0.769 (0.032)
US Female	0.065 (0.052)	0.198 (0.043)	0.691 (0.061)	0.803 (0.064)
Non-US Female	-0.187 (0.028)	-0.183 (0.027)	0.389 (0.045)	0.66 (0.055)
Observations	6,320	6,320	1,309	1,309

We can gain further insight by looking at how admissions vary with applicant quality and identity.[17] Column 1 in table 24 illustrates the effect observable in the descriptive statistics: that the probability of acceptance is lower for non-US applicants. The question, then, is whether this is due to their identity or their other characteristics. Columns 2 and 3 report the relationship between individual identity and subjective performance scores. We see here that non-US applicants still have a lower probability of acceptance, though the effect of identity is much reduced. The final column includes both scores. In this case, there is little evidence that identity affects admissions. There is a slightly positive relationship between being a US female applicant and the probability of admission to a top-ten program. Moreover, the results imply that the subjective and objective scoresmeasure different attributes. It is also the case that the objective score is more important for non-US applicants, consistent with the hypothesis that these scores are important for identifying skilled individuals from different regions.

We can also examine admission probabilities along the distribution of applicants within each subgroup. For individual i, let $q_i \in [0,1]$ be the person rank; that is q_i fraction of applicants have a probability of admissions that is lower than p_i, individual $i's$ estimated probability of admission to a top-ten program. More precisely, given an individual $i \in I^g$, with identity $g \in \{USMale, USFemale, NonUSMale, NonUSFemale\}$, we can compute the number of persons in group g with rank less than q:

$$n_L^g(q) = \#\{q_j \leq q | j \in I^g\}.$$

We can also compute the number of individuals in this group who are admitted to a PhD program:

$$n_{LA}^g(q) = \#\{q_j \leq q | i \in I^g, a_j = 1\},$$

where a_j is 1 if and only if individual j is admitted to a top-ten program. Given these numbers, for each person $i \in I^g$ we can compute the probability that a person in their group with a lower rank

[17] There is a large literature on the effect of identity on wage inequality by income quantile. See Fortin et al. (2010) for a comprehensive review of decomposition methods.

TABLE 24. Identity and Scores in Top 10 Admissions

	T10			
	(1)	(2)	(3)	(4)
US Female	0.121	0.323	0.078	0.237
	(0.104)	(0.103)	(0.094)	(0.103)
Non-US Female	−0.682	−0.270	−0.474	−0.158
	(0.061)	(0.072)	(0.082)	(0.097)
Non-US Male	−0.333	0.031	−0.267	0.036
	(0.104)	(0.140)	(0.087)	(0.128)
Objective Score		1.005		0.781
		(0.083)		(0.097)
Subjective Score			1.253	1.120
			(0.059)	(0.058)
Constant	−1.045	−1.566	−1.500	−1.879
	(0.085)	(0.133)	(0.094)	(0.140)
Pseudo R^2	0.0105	0.0951	0.1709	0.2114
Node Purity	0.6037	0.6303	0.6551	0.6688
Observations	6,320	6,320	6,320	6,320

Note. Standard errors are clustered by application year and are heteroskedasticity-robust (HC3).

is admitted to a top-ten PhD program:

$$P_L^g(q_i) = \frac{n_{LA}^g(q)}{n_L^g(q)}.$$

This number is always well-defined and provides a measure of how the probability of low-ranked individuals increases with rank. Also, the number of individuals in our sample is sufficiently large that we do not need to smooth the results curves. The results are illustrated in figure 9. If we were able to predict admissions perfectly, then we would have $P_L^{males}(q) = 0$ for $q \leq 1 - 0.207 = 0.793$,and then a diagonal line to 0.207. This is illustrated by the red line labeled "Perfect Selection". As one can see, most of the plotted curves are above this, illustrating that many lower-ranked individuals are admitted (where rank is determined by our model). Notice that there are non-US females with low scores but a higher probability of being admitted than US males. However this trend reverses after the midpoint, and non-US females are admitted with a lower probability than US males for $q_i > X$. For high enough quantiles, US applicants tend to be admitted at a higher rate than non-US applicants. While these results suggest that our measures provide good controls, the overall fit of the model has an R^2 of 0.22 and a node purity of 0.66 (a node purity of 0.5 would indicate random choice). Thus, there is evidence of quite a bit of noise in the system. The next section explores the reliability of the subjective performance measures as a potential source of noise.

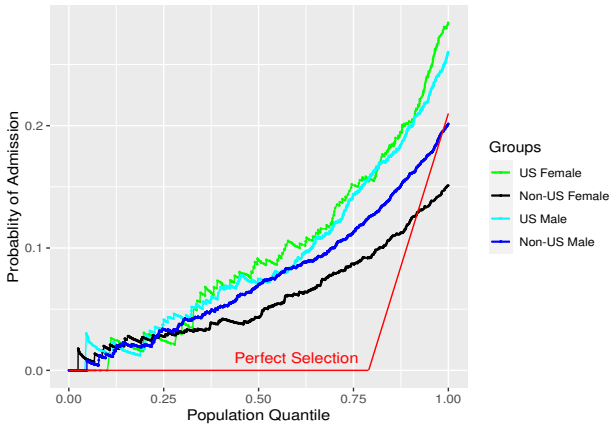


FIGURE 9. Probability of Admissions Given Score Is Less Than Population Quantile

A.9. **Reliability of Subjective Ratings.** This subsection replicates the analysis in section 7 but only for the subjective ratings submitted by references.

We first present some descriptive statistics regarding the subjective ratings. Table 25 presents the most frequent pattern ranked by aggregate score. Column 1 is the percentile ranking of the subjective ratings score θ^s using the weights reported in table 21. The responses are in the rating columns, where AP is academic potential, IP is intellectual performance, RP is research potential, and WS is writing skill. Interestingly, the vector of ratings with the largest positive effect on admissions has NA for the research potential response. This illustrates that all signals, including non-response, may be used to make inferences regarding quality.

The third entry corresponds to straight 1s, the highest possible rating. This is by far the most frequent response, corresponding to 24% of all applications and 40% of admitted students. Note that the next most frequent combination is 1s for AP, IP, and RP, but a 2 for WS. This combination is observed in almost 9% of the cases, and 12.5% of the admitted students. It is clear that high ratings are associated with higher admissions probabilities. The top rating is supposed to be given only to "exceptional" students well above the 5% quality. In particular, given that 20% of the sample is admitted to top-ten programs, one might expect a 100% acceptance rate for this group. That this is not the case implies that committees do not rely completely on these ratings, suggesting a concern around upward bias. If all references for an applicant give the highest rating, this would be expected to increase a committee's confidence in the quality of the applicant. As discussed in the theory section, another source of information is to use the comparative ratings from the same reference for different applicants.

TABLE 25. Subjective Ratings Response Patterns With More Than 100 counts, and the Max and Min Weights

Percentile	Rating				Number of Letters		Fraction of Letters (%)	Fraction of Letters for Admitted (%)
of Mean Rating	AP	IP	RP	WS	Total	T10 Admits	Total 14,598	Total 2,889
1.000	1	1	NA	1	5	1	0.030	0.030
1.000	1	2	3	1	1	0	0.010	0
0.874	1	1	1	1	3,621	1,164	24.800	40.290
0.689	1	2	1	2	136	23	0.930	0.800
0.689	1	1	1	2	1,288	363	8.820	12.560
0.689	1	1	1	Unknown	192	48	1.320	1.660
0.617	1	1	1	3	197	52	1.350	1.800
0.599	1	2	2	2	139	23	0.950	0.800
0.599	1	1	2	2	139	24	0.950	0.830
0.573	2	1	1	1	334	71	2.290	2.460
0.532	NA	NA	NA	NA	696	141	4.770	4.880
0.478	2	2	1	2	263	41	1.800	1.420
0.478	2	1	1	2	492	81	3.370	2.800
0.427	2	2	2	1	140	18	0.960	0.620
0.324	2	2	2	2	1,840	300	12.600	10.380
0.324	2	2	2	Unknown	165	29	1.130	1
0.324	2	1	2	2	113	15	0.770	0.520
0.181	2	2	2	3	407	42	2.790	1.450
0.152	3	2	2	2	153	13	1.050	0.450
0.119	3	3	3	3	547	56	3.750	1.940
0.080	3	2	2	3	154	9	1.050	0.310
0.042	3	3	3	4	109	7	0.750	0.240
0.011	4	4	4	4	182	9	1.250	0.310
0.0003	5	4	4	5	9	0	0.060	0

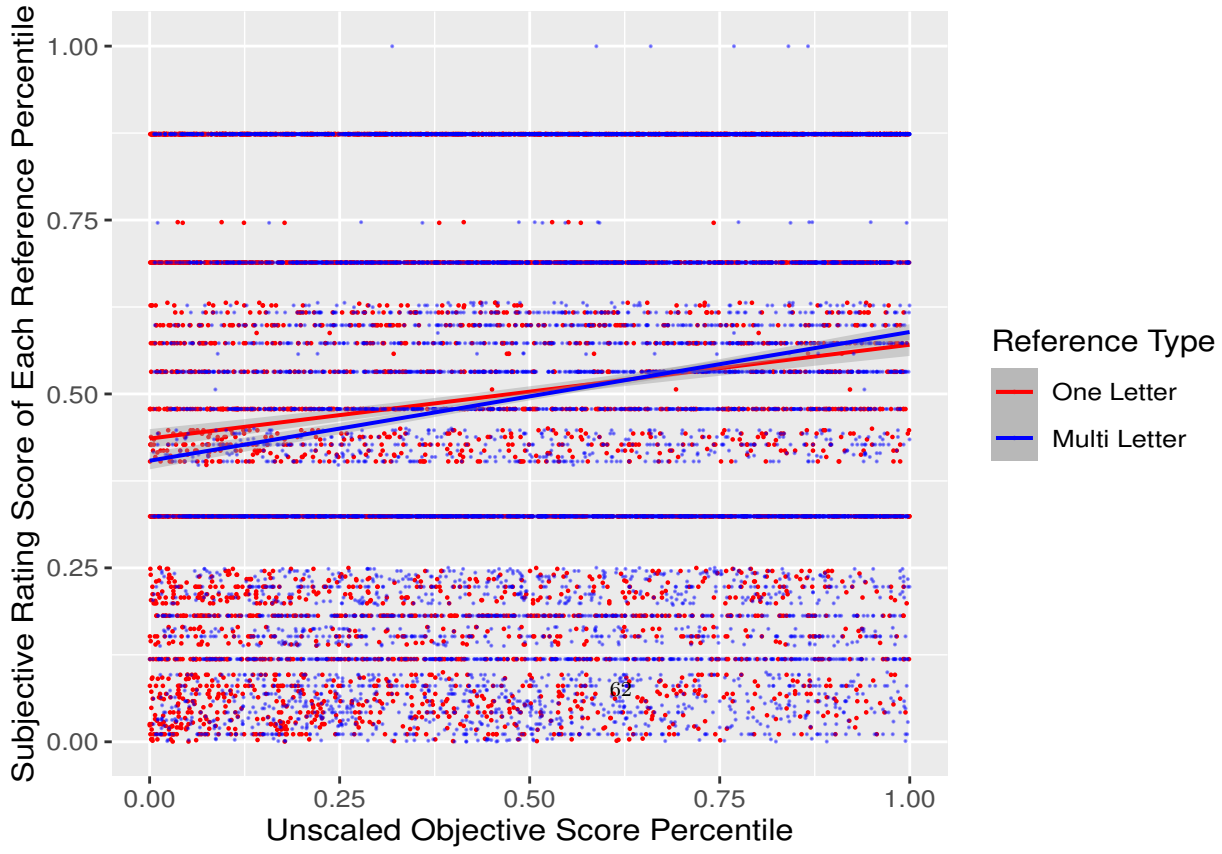


TABLE 26. Correlation Between Objective and Individual Level Subjective Rating

	Dependent variable: Subjective Score Percentile	
	SingleRef	MultiRef
	(1)	(2)
Objective Score Percentile	0.135	0.185
	(0.013)	(0.011)
Constant	0.436	0.404
	(0.007)	(0.006)
Observations	5,283	9,315
Adjusted R^2	0.020	0.032

Note. Standard errors are heteroskedasticity-robust (HC3).

FIGURE 11. Individual Subjective Rating versus Applicant Mean Rating

A.10. **Job Market Success by GRE Quantitative Scores.** Table (28) reports the fraction of the pool by GRE ventile (except for those in 0-79, who are pooled). The "Admitted" column is the fraction of the individuals in our data in that GRE ventile who are admitted to any program. The "Other Programs" are individuals who are admitted to some PhD program outside of the top ten. The "Top 10 AP" and "Top 20 AP" are the fractions of the *admitted* students in the reported

TABLE 27. Correlation Between Average Subjective Rating and Individual Rating

	Dependent variable: Subjective Rating Percentile	
	SingleRef	MultiRef
	(1)	(2)
Average Subjective Rating	0.961	1.026
	(0.017)	(0.013)
Constant	0.027	−0.017
	(0.011)	(0.008)
Observations	5,283	9,315
Adjusted R^2	0.475	0.499

Note. Standard errors are heteroskedasticity-robust (HC3).

GRE range who obtained a top 10 or 20 assistant professor position. The 14% in the top 20 AP for top-ten programs corresponds to a single person. Note as well the non-linearity above the 85 percentile for the top ten admits. This is consistent with the hypothesis that above the 90 percentile the GRE is not predictive of placement.

TABLE 28. Success Rates By Quantitative GRE Scores

	Top 10 Programs			Other Programs		
GRE Q	Admitted	Top 10 AP	Top 20 AP	Admitted	Top 10 AP	Top 20 AP
0-79	5%	0.00%	14%	21%	0.00%	0.00%
80-84	3%	0.00%	0.00%	32%	0.00%	5%
85-89	9%	4%	9%	39%	2%	3%
90-94	19%	11%	18%	36%	1%	1%
95-99	31%	6%	10%	34%	0.3%	1%

HARVARD UNIVERSITY, DEPARTMENT OF ECONOMICS, CAMBRIDGE, MA

COLUMBIA UNIVERSITY, DEPARTMENT OF ECONOMICS, NEW YORK, NY

COLUMBIA UNIVERSITY, DEPARTMENT OF ECONOMICS, NEW YORK, NY

COLUMBIA UNIVERSITY, DEPARTMENT OF ECONOMICS, NEW YORK, NY