

ARE UNIVERSITY ADMISSIONS ACADEMICALLY FAIR?

Debopam Bhattacharya, Shin Kanaya, and Margaret Stevens*

Abstract—Admission practices at high-profile universities are often criticized for undermining academic merit. Popular tests for detecting such biases suffer from omitted characteristic bias. We develop a bounds-based test to circumvent this problem. We assume that students who are better qualified on observables would, on average, appear academically stronger to admission officers based on unobservables. This assumption reveals the sign of differences in admission standards across demographic groups that are robust to omitted characteristics. Applying our methods to admissions data from a British university, we find higher admission standards for men and slightly higher ones for private school applicants, despite equal admission success probability across gender and school background.

I. Introduction

ADMISSION practices at selective universities generate considerable public interest and political controversy due to their close connection with intergenerational mobility and social discrimination. For example, in the United Kingdom, a highly publicized 2011 Sutton Trust report shows that nationally just 3% of schools—mostly expensive and independent (as opposed to state-run) institutions—account for 32% of undergraduate admissions to Oxford and Cambridge, while these universities claim to admit solely on the basis of academic merit. Background-based admission quotas such as caste-based reservation in India and race-based affirmative action in the United States have generated intense public controversy. Despite significant public interest in these issues, rigorous methods for modeling and testing the “fairness” of admissions based on empirical evidence are absent in the academic literature. In this paper, we develop an empirical framework to model meritocracy of admission decisions and use it to infer whether all applicants are held to the same academic standard during admissions.

A simple approach to detecting discrimination in admissions, popular in the education literature, is to test if demographic or socioeconomic characteristics of applicants are significant determinants of admission, after controlling for commonly observed academic records such as past test scores (Espenshade, Chung, & Walling, 2004; Zimdars & Heath, 2009; Hurwitz, 2011). However, if admission officers observe more indices of academic ability than the researcher, and the relation between observable and unobservable indices varies by demographics, then these naive tests become invalid (Heckman, 1998). For instance, if female candidates *ceteris paribus* perform better on interviews and interview scores are unobserved by a researcher, then an equal admission rate of observationally similar male and female candidates implies bias against female applicants. Indeed, in the empirical context investigated in this

paper, we find that socioeconomic backgrounds do not have statistically significant effects on admission rates once we control for preadmission test and interview scores. However, applying a more careful analysis that addresses the omitted characteristic problem, we find that male candidates face a higher admission threshold than female candidates and that differences in thresholds across type of school attended by the applicant are less significant.

Beyond their obvious legal and political significance, such findings also have important policy implications. For example, knowing that one has to admit weaker female students to maintain gender balance in application success rates raises questions about what investments are needed at the school level to improve the quality of female applicants. Naive satisfaction with gender equality in admission success would conceal this important role for potential interventions.

Methodologically, our approach to bias detection is related to the productivity-based view of optimal decisions in the tradition of Becker (1957). Viewed in this light, if admissions are purely meritocratic, then the marginal admitted student from a state school should be expected to perform equally well in postadmission assessments (e.g., college exams) as the marginal student admitted from a private school, but her expected performance would be worse under affirmative action. Conversely, taste-based discrimination against state schools will lead to the marginal state school admitted student to perform better than the marginal candidate admitted from an independent school. The difference between expected performances of marginal candidates across demographic groups can therefore be interpreted as a measure of deviation from meritocracy.

A challenge in implementing this approach directly is that a researcher typically observes a subset of the applicant characteristics used by admissions officers, and the distributions of the unobserved characteristics may, and usually do, differ across demographic groups. This “omitted characteristics” problem jeopardizes the researcher’s attempt at reconstructing the decision maker’s perceptions and identifying the marginal admits, and therefore assessing whether the decision maker acted in an academically unbiased way. Problems of this type have been recognized by previous researchers in the context of detecting taste-based discrimination in hiring (Heckman, 1998). In this paper, we devise a test for meritocratic admissions—based on the differences in admission thresholds faced by different demographic groups—that, under appropriate assumptions, is robust to the omitted characteristics problem.

Specifically, we construct an empirical, threshold-crossing model of admissions involving observed applicant covariates and unobserved heterogeneity—applicant characteristics observed by admission officers, but unobserved by the researcher. In our model, academic fairness corresponds to

Received for publication December 11, 2014. Revision accepted for publication April 4, 2016. Editor: Asim I. Khwaja.

* Bhattacharya: University of Cambridge; Kanaya: University of Aarhus; Stevens: University of Oxford.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/REST_a_00618.

using identical thresholds of expected future performance across applicants from different demographic groups. Our key assumption, for which we will provide supporting empirical evidence, is that students who are significantly better in terms of easily observable indicators of academic potential should statistically—but not necessarily with certainty—be more likely to appear stronger to the admission officer, based on characteristics observed by her but not by the researcher. The distribution of unobservables, conditional on observables, is otherwise allowed to be arbitrarily different across demographic groups. We show that by using this assumption in conjunction with pre- and postenrollment data, one can learn about the sign of the differences between admission thresholds applied to different demographic groups.

We use our methods to analyze admissions data from a selective U.K. university on applicants who have cleared an initial exam-based elimination round. We first provide evidence in support of our identifying assumption; we then apply our methods to show that male applicants face a higher admission standard than females,¹ whereas standards faced by private school applicants are possibly slightly higher than those faced by state school applicants. In contrast, the application success rates are very similar across gender and type of school attended by the candidate, both before and after controlling for key covariates, thereby illustrating the crux of our approach.

A large volume of research exists in educational statistics on the analysis of admissions to selective colleges, focusing mainly on the United States (Hoxby, 2009). In this context, our goal is to assess the extent of meritocracy in prevalent admission practice by focusing on the marginal admits in different demographic groups. This enables us to demonstrate empirically that equal success rates in admissions across demographic groups can be consistent with very different admission standards across these different groups. (See Sander, 2004, for an early discussion of these issues in the context of U.S. law school admissions.) This is in contrast to many other studies, both academic and policy oriented, that compare either average preadmission test scores (Herrnstein & Murray, 1994) or average postadmission performance across all (as opposed to marginal) admitted students from different socioeconomic groups (Keith et al., 1985; Waters et al., 2009; Kane, 1998).

Our paper also complements an existing literature on analyzing the consequences of affirmative actions in college admissions. Fryer and Loury (2005) provide a critical review of the relevant theoretical literature. On the empirical side, Arcidiacono (2005) uses a structural model of admissions to simulate the potential counterfactual consequences of removing affirmative action in U.S. college admission; Card and Krueger (2005) describe the reduced-form impact

of eliminating affirmative action on minority students' application behavior in California, and Hinrichs (2012) examines the effects of banning preferential admission policies on enrollment patterns of both minority and nonminority students. Arcidiacono and Lovenheim (2015) provide a review of the empirical evidence on the effect of affirmative action on student-college mismatch. This paper, though substantively related to the work we identified, has a different goal. Here, we construct a formal econometric model where affirmative action (or taste-based discrimination) and meritocracy have different empirical implications, and we use it in conjunction with admissions-related microdata to detect deviations from meritocracy. To our knowledge, the only other work in this literature that focuses on marginal admits is Bertrand, Hanna, and Mullainathan (2010), who examined the consequences of affirmative action in admission to an Indian college. In their setting, admission was based on score in a single entrance exam; admission thresholds differed by applicants' social caste and were publicly announced. This setup removes a key empirical challenge—that of defining and identifying the marginal admits and rejects—arising in general admissions contexts where entrance is based on several background variables, there is unobserved heterogeneity across applicants, and admission thresholds are not explicitly announced. Our context requires us to deal with this more general scenario.

Although this paper focuses on the issue of college admissions, the general methodology is applicable to many other settings of testing bias in institutional decision making. Common examples include approval of business loan and mortgage applications, referrals to expensive surgery versus cheaper medicine-based treatment, and hiring decisions. The data setting is one where a researcher has access to key characteristics of individual applicants and the eventual decision made on their behalf by the approval agency. These key characteristics need not be exhaustive, and this paper's methodology allows the possibility that approvers may observe a richer set of applicant characteristics than the researcher does. Applying our methods, one can then test whether the observed data are consistent with meritocratic approval processes, for example, that all loan applicants face a common ceiling of default probability below which the application is approved, or that each patient has to clear the same hurdle of expected survival days following the surgery in order to qualify for the procedure.

The rest of the paper is organized as follows. Section II sets up a simple theoretical model, followed by the corresponding empirical model of meritocratic admissions. Section III describes the data. Section IV states the assumptions, provides empirical evidence in support of the key identifying assumption, and lays out the identification analysis. Section V discusses inference. Section VI reports the empirical findings from the real data set, presents robustness checks, and discusses some caveats. Section VII concludes. An online appendix contains the basic economic model of optimal admissions (part A), some additional figures and

¹ As a referee has pointed out, it remains possible that some academically stronger female candidates were erroneously eliminated in the first round; had they been retained, the gender gap may have appeared narrower.

tables relevant to robustness checks (B.1 and B.2), the result of a simulation exercise based on the real data (part B.3), and formulas for calculating the confidence intervals for threshold differences (part C).

II. Benchmark Optimization Model

In the online appendix, part A, we lay out a benchmark economic model of admissions to help fix ideas. Based on this economic model, we develop a corresponding econometric model incorporating unobserved heterogeneity, which can be used for admissions data. The basic elements of the economic model are as follows.

Let W denote an applicant's preadmission characteristics, observed by the university. Let $\phi(w)$ denote a w -type student's expected outcome (e.g., expected future GPA) if he or she enrolls, and let $\alpha(w)$ denote the probability that a w -type student, on being offered admission, eventually enrolls. Let $c \in (0, 1)$ be the fraction of applicants who can be admitted, given the number of available spaces. If the university wishes to maximize total performance of the incoming cohort subject to the restriction on the number of vacant places, then its admission strategy would be to admit individuals whose $\phi(w) \geq \gamma$, where γ is chosen to satisfy the budget constraint. The key feature of this rule is that γ does not depend on covariates, and so the value of an applicant's W affects the decision on his or her application only through its effect on $\phi(W)$. To get some intuition on this, consider the case where one of the covariates in W is gender, and assume that the admission threshold for women, γ_{female} , is strictly lower than that for men, γ_{male} . Then the marginal female, admitted with $w = (x, female)$, contributes $\gamma_{female} \times \alpha(x, female)$ to the expected aggregate outcome and takes up $\alpha(x, female)$ places, implying a contribution of γ_{female} ($= \alpha(x, female) \gamma_{female} / \alpha(x, female)$) to the objective of average realized outcome. Similarly, the marginal rejected male, if admitted, would contribute γ_{male} to the average outcome. Since $\gamma_{male} > \gamma_{female}$, we can increase the average outcome if we replace the marginal female admit with the marginal male reject. Thus, different thresholds cannot be consistent with the objective of maximizing the overall outcome. Our goal is to use actual admissions data to understand whether admission officers use identical thresholds across sociodemographic groups. The key challenge is to allow for the possibility that admission-officers' inference about academic merit were based on more characteristics than the researchers observe, so that one cannot infer the admission thresholds simply based on observed characteristics. We now turn to the task of constructing an econometric model incorporating unobserved heterogeneity in an empirical model of admissions.

A. Econometric Model

To set up the empirical framework, let $W := (X, G)$, where G denotes one or more discrete components of W

capturing the group identity of the applicant (such as sex, race, or type of high school attended), which forms the basis of commonly alleged mistreatment. The variables in X are the applicant's other characteristics observed prior to admission, which include one or more continuously distributed components like standardized test scores. We observe the covariates X, G and the binary admission outcome D ($= 1$ if admitted and $= 0$ otherwise). Let $\mathcal{X}_g, \mathcal{X}_h$ denote the support of X for applicants of type $G = g$ and $G = h$, respectively.

Let Z denote an index of academic ability of applicants, based on "soft" characteristics, such as evidence of enthusiasm and academic reference letters, which are unobservable to the analyst but observed by the admission officer. This may also include any random idiosyncrasies in the officers' expectation formation process.² We assume that larger values of Z , without loss of generality, denote higher perceived academic potential.

Under meritocratic admissions, admission officers would decide on whether to admit applicant i in the current year, based on $\phi(X_i, G_i, Z_i)$, their subjective assessment of i 's academic merit (e.g., how applicant i will perform when admitted).³ In accordance with our economic model, we assume that applicant i with $G_i = g$, $Z_i = z$, and $X_i = x \in \mathcal{X}_g$ is offered admission (i.e., $D_i = 1$) if and only if $\phi(x, g, z) \geq \gamma$, where γ denotes the university-wide baseline threshold for applicants—that is,

$$D_i = \begin{cases} 1 & \text{if } \phi(X_i, G_i, Z_i) \geq \gamma; \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

An admission practice is academically fair if and only if γ does not vary by demographics. The underlying intuition is that the only way covariates G should influence the admission process is through their effect on the perceived academic merit. Having a larger γ for, say, women than men implies that a male applicant with the same expected outcome as a female applicant is more likely to be admitted. Conversely, under affirmative action policies, γ will be lower for demographics that represent historically disadvantaged groups. Therefore, we are interested in testing whether the values of the threshold γ are identical across demographics. We will call γ the admission threshold.

Thus, a female applicant with identical X as a male candidate can have a higher probability of being admitted, and yet the admission process may be academically fair if women have a higher expected performance than men with identical X . This notion of fairness differs from one that requires that individuals who are identical on publicly verifiable variables (i.e., the X s) must have equal chances of getting in, no matter what their value of G and no matter whether predicted future performance differs across G for the same value of X .

² When there are multiple sources of soft information, Z may be interpreted as a composite scalar index (e.g., a weighted average) of these characteristics.

³ In line with the existing literature on bias detection referenced above, we ignore issues about risk and leave that for future research.

Remark 1. It is important to note that we do not assume that admission officers literally calculate expected future performance in order to admit candidates. Our goal is to assess whether the admission process, whatever its goal and how-ever it is conducted, is consistent with the goal of admitting the academically strongest applicants.

III. Data

Our empirical analysis is based on admissions data for two recent cohorts of applicants to a competitive and popular undergraduate degree program at a selective U.K. university. Students enter British universities to study a specific subject, rather than the U.S. model of starting a general curriculum, followed by specialization in later years. Consequently, admissions are conducted primarily by faculty members (referred to as admission tutors) in the specific discipline to which the candidate has applied. An applicant competes with all others who apply to this specific subject, and no switches are permitted across disciplines in later years. The admission process is held to be strictly academic; extracurricular achievements are given no weight. In that sense, these admissions are more comparable to Ph.D. admissions in U.S. universities. Furthermore, almost all U.K. applicants take two common school-leaving examinations, the General Certificate of Secondary Education (GCSE) and the A-levels, before entering university. Each of these examinations requires the student to take written tests in specific subjects. The examinations are centrally conducted, so the scores of individual students are directly comparable. In addition, all applicants take a multiple-choice aptitude test, similar to the SAT in the United States, and write an essay that is graded.

A. Choice of Sample

For our empirical analysis, we focus on U.K.-domiciled applicants. The application process consists of an initial stage whereby a standardized form, the Universities and Colleges Admissions Service (UCAS) is filled by the applicant and submitted to the university. This form contains the applicant's unique identifier number, gender, school type, prior academic performance record, personal statement, and a letter of reference from the school. The GCSE, the aptitude test, and essay scores are separately recorded. About one-third of all applicants are then selected for interview by admission tutors on the basis of the aptitude test. The rest are rejected. Selected candidates are then assessed in a face-to-face interview, and the interview scores are recorded centrally. This subgroup of applicants who have been called to interview will constitute our sample of interest. Therefore, we are in effect testing the academic efficiency of the second round of the selection process, taking the first round as given. Accordingly, from now on, we will refer to those summoned for interview as the applicants. The final admission decision is made by considering all candidate-specific information from

TABLE 1.—VARIABLE LABEL

gcsescore	Overall score in GCSE, 0–4
alevelscore	Average A-level scores, 80–120
aptitude test	Overall score in aptitude test, 0–100
essay	Score on substantive essay, 0–100
interview	Performance score in interview, 0–100
prelim_avg	Average score in first year university exam, 0–100
offer	Whether offered admission

The gcsescore is an average of the GCSE grades achieved by the candidate for eight subjects, where A* = 4, A = 3, B = 2, C = 1, D or below = 0. The grades used are mathematics plus the other seven best grades. The alevelscore is an average of the A-levels achieved by or predicted for the candidate by his or her school, excluding general studies. Scores are calculated on the scale A = 120, A/B = 113, B/A = 107, B = 100, C = 80, D = 60, E = 40, as per England-wide UCAS norm.

among the applicants called for interviews. For our application, we use anonymized data for two cohorts of applicants from their records held at the central admissions database at the university. To preserve anonymity, the data do not contain reference letters.

B. Choice of Covariates

We chose a preliminary set of potential covariates to be the observables, based on the information recorded on UCAS forms and the university's application records. We use as observable components (i.e., X) GCSE score, aptitude test scores, the examination essay score, and the interview score. A more detailed description of these covariates is provided in table 1. The unobservable index of achievement Z pertains to information conveyed by recommendation letters. Given that those summoned for interview constitute our population of interest, we found that in terms of whether the applicant previously read two subjects recommended for entry, there is very little variation across these applicants and including these covariates makes no difference to our eventual results. Therefore, we eventually dropped these variables from the analysis.

C. Group Identities G

We consider the academic efficiency of admissions with regard to two different group identities: type of school the applicant attended and the applicant's gender. Selective universities in the United Kingdom are frequently criticized for the relatively high proportion of privately educated students admitted. The implication is that applicants from independent schools, where spending per student is much higher than in state schools (Graddy & Stevens, 2005), have an unfair advantage in the admission process. This is of special concern in a country like the United Kingdom where most selective universities are largely funded by taxpayers. The issue of gender differences in admission and academic performance is, of course, a more universal issue. In the United Kingdom, as in most other OECD countries, the higher education participation rate is higher for women, having overtaken that for men in 1993. However, selective universities in the United Kingdom have lagged behind the trend: in 2010–11, 55% of undergraduates across all U.K. universities were women, but 44% of students admitted to

the university we are analyzing were women. Typically, gender imbalances are more pronounced in certain programs and include the one we study, where male enrollment is nearly twice the female enrollment.

In our data set, we can also match the postadmission academic performance of admitted students to their preadmission characteristics. In principle, one can use this information for analyzing potential bias in admissions. Allowing for selection on unobservables, however, means that such data cannot be used without making more restrictive assumptions. For example, a regression of eventual academic performance on preadmission covariates for admitted candidates does not yield a consistent estimate of the predictive power of these covariates for the pool of applicants, for whom the admission decision is made. Indeed, due to classical selection bias, one would expect such effects to be biased toward 0 (see Rothstein, 2004, for a discussion of related issues). A second potential limitation of such data is that academic performance as measured by the university's own exams may not be the sole index of academic ability sought by an admission tutor. They might focus instead on a subjective measure of academic ability that may be positively correlated only with eventual performance on university exams. For these reasons, we did not include these data in our main analysis. Nonetheless, while interpreting our empirical results, we use these predictive regressions (see figures 3 and 6) as suggestive evidence of where our results might have arisen from.

IV. Assumptions

In order to develop a test of meritocratic admissions that can be applied to the above data, we will make a set of assumptions using the following notation. For any pair of individuals i and j , where i is of type g and has a value of X equal to x_g and j is of type h and has $X = x_h$ with $x_g \in \mathcal{X}_g$ and $x_h \in \mathcal{X}_h$, the notation $x_g \succeq_\varepsilon x_h$ will mean that applicants i and j are identical with respect to all qualitative attributes and, moreover, every continuously distributed component of x_g is at least ε standard deviations larger than the corresponding component of x_h . For example, if $G = \text{school type}$ and $X = (\text{SAT}, \text{GPA}, \text{male})$, then $x_g \succeq_\varepsilon x_h$ means that applicant i and j are both male or both female and that $\text{SAT}_i > \text{SAT}_j + \varepsilon\sigma_{\text{SAT}}$ and $\text{GPA}_i > \text{GPA}_j + \varepsilon\sigma_{\text{GPA}}$, where σ_{GPA} and σ_{SAT} are the standard deviation of GPA and SAT for the entire population of applicants. We will denote by $Q^\tau(Z|A)$ the τ th quantile of the random variable Z given the random variable A .

Throughout the rest of the paper, we maintain the following assumption:

Assumption M (median restriction). (i) There exists $\varepsilon > 0$ such that for any $e \geq \varepsilon$, if $x_g \in \mathcal{X}_g$ and $x_h \in \mathcal{X}_h$ and $x_g \succeq_e x_h$, then,

$$\begin{aligned} \text{Median}[Z|X = x_g, G = g] \\ \geq \text{Median}[Z|X = x_h, G = h], \end{aligned}$$

for any g and h ; (ii) $\phi(X_i, G_i, Z_i)$, introduced just before equation (1), is continuously distributed conditionally on any realization of (X_i, G_i) .

A stronger version of assumption M is first-order stochastic dominance, which has the same intuitive interpretation as assumption M:

Assumption SD (stochastic dominance). There exists $\varepsilon > 0$ such that for any $e \geq \varepsilon$, if $x_g \in \mathcal{X}_g$ and $x_h \in \mathcal{X}_h$ with $x_g \succeq_e x_h$, then the distribution of Z conditional on $X = x_g, G = g$ first-order stochastically dominates that of Z conditional on $X = x_h, G = h$,

$$\Pr[Z \leq a|X = x_g, G = g] \leq \Pr[Z \leq a|X = x_h, G = h],$$

for any a and for all g, h ; (ii) $\phi(X_i, G_i, Z_i)$ is continuously distributed conditionally on any realization of (X_i, G_i) .

Crudely speaking, assumption M/SD means that applicants who are better along standard, observable indicators of academic ability are also likely to be better “on average” in terms of the index of unobserved characteristics that the tutors weigh positively in determining admissions. The motivation for this assumption comes from the fact that for meritocratic admissions, the outcome of interest may be thought of as a measure of future academic performance, whereas the measures in X are a set of past academic performance in high school or admissions-related assessments. It is therefore likely that candidates who have performed significantly better in past assessments are statistically more likely to have performed better in those assessments (unobserved by the researcher) that admission tutors view as positive determinants of future performance and, hence, under the assumption of being academically motivated, would weigh positively on the decision to admit. While assumption M/SD is likely to hold for the population of all students, some of this positive dependence may be partially eroded for the population of applicants if the decision to apply depends on unobservables. Indeed, if applications are costly and a student applies despite having low scores on observable tests, she is likely to be stronger on unobservable attributes relative to the average student with low observable test scores in the population. Such selective application will reduce the extent of positive dependence between observables and unobservables among the applicants relative to that in the population of all students. We address this concern by providing evidence strongly suggesting that the aggregate impact of such “erosion” on the positive dependence is likely insignificant.

The magnitude of ε controls the strength of assumption M. Thus, $\varepsilon = 0$ corresponds to the benchmark case where we are comparing a pair of g - and h -type applicants, such that the former has scored higher in each previous assessment than the latter. A strictly positive ε leads to comparison of applicant pairs with no overlap of preadmission test scores. The higher is ε , the more likely are assumptions M or SD

to hold, but the lower will be the power of our test, since fewer pairs of students will satisfy M/SD with a higher ϵ . A practical method for choosing ϵ in an application is suggested below.

Note also that assumption M is substantively much weaker than two informal arguments often used in applied work: (a) when the distribution of the observable covariates is balanced across treatment and control groups in quasi-experimental designs, it is taken to imply that they are also balanced in terms of unobservables (Greenstone & Gayer, 2009), and (b) orthogonality of an instrument with observed covariates is taken as suggestive evidence that it is orthogonal with unobserved covariates (Angrist & Evans, 1998). In our context, the types of variables typically unobservable to researchers but likely to affect admissions include achievements such as winning special academic prizes, participation in science or math olympiads, high intellectual enthusiasm conveyed by applicants' personal essays, and the subjective impressions of previous teachers implied in recommendation letters. Such specific information can identify individual applicants and therefore are most likely to be withheld from researchers owing to privacy considerations. However, while making admission decisions, tutors are likely to observe these characteristics for current applicants through their application materials or through personal interactions. It is intuitive that such achievements are statistically more likely to have occurred for individuals who score higher in terms of easily observable entrance assessments and aptitude tests than those who score lower.

Finally, the continuity condition in assumption $M(ii)$ rules out gaps in the distribution of Z , which helps to relate the probability of admission to the admission thresholds. Such continuity is intuitive, especially when Z is a function of several underlying performance indicators that are themselves continuously distributed.

Remark 2. Note that assumption M/SD does not say that applicants with higher X have higher Z with probability 1; it simply says that their values of Z tend to be higher in a stochastic sense.

Remark 3. The restriction on the median cannot be replaced by a restriction on the conditional expectation for identification purpose since we are considering a discrete choice problem: $D = \mathbf{1}\{\phi(X, G, Z) \geq \gamma_G\}$. See Manski (1988) for why a conditional quantile restriction is necessary for the identification of discrete-choice models.

Remark 4. Assumption M allows the distribution of the unobservable Z to differ by background variables. In particular, we allow both the location and the scale of Z to depend on G (conditional on X) and thus also allow for the realistic situation of larger uncertainty regarding applicants from historically underrepresented communities.

A. Empirical Evidence of Median Dominance

Among the preadmission variables that we observe in our data set, only the score on the interview is assigned by tutors. This is the type of variable most likely to be missing in other data sets since they reflect subjective assessment by the admission tutors. We will first check our assumption M for the applicants in our data by treating the interview score as the unobservable component. That is, we will verify whether the median interview score is higher for those types of applicants who are better in terms of all other "tutor-independent" test scores X obtained in prior assessments. If applications are costly, a student with low scores on X will apply only if her potential performance on the interview is likely to be high, so that an applicant with low X is likely to be stronger on interview skills relative to the average student with low X . The question is whether this negative relationship is strong enough to override the overall positive relationship in the population. Since the interview score is observed for the entire sample, we can test this hypothesis.⁴ The concrete steps leading to our test are as follows. Consider $X = (\text{GCSEscore}, \text{Aptitude_test_score}, \text{Exam_essay})$. First, run a median regression of interview score (which now plays the role of Z) on X and quadratics in components of X plus G , where G represents gender or school type, and compute the predicted values. These represent $\text{Median}[Z|X, G]$. We then compare these predicted values for pairs of applicants where the first applicant is of type $G = g$ and the second applicant is of type $G = h$. In figure 1, we depict histograms capturing the marginal distribution of the conditional median differences for different combinations of g and h . The analog of our assumption M here is that these histograms should have an entirely positive support, up to estimation error. For example, the histogram in the top-left panel of figure 1 shows the estimated marginal distribution of the variable,

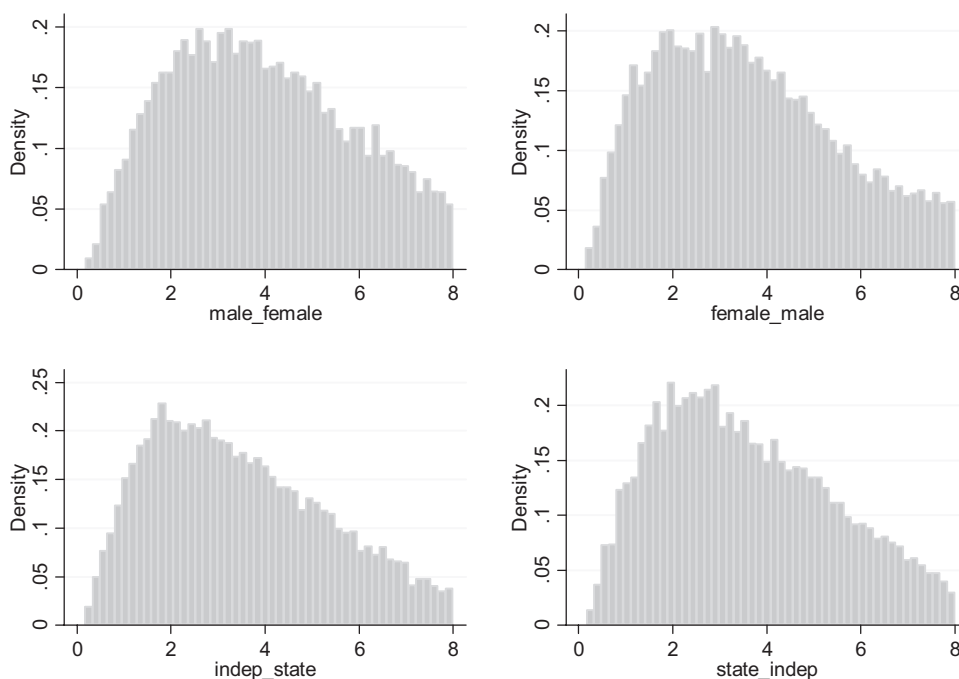
$$\begin{aligned} & \text{Median}[\text{interview} | X_g, g = \text{male}] \\ & - \text{Median}[\text{interview} | X_h, h = \text{female}], \end{aligned}$$

across all paired realizations (X_g, X_h) satisfying $X_g \geq_\epsilon X_h$. We choose $\epsilon = 0.0$. If we demonstrate median dominance for $\epsilon = 0.0$, then dominance will obviously hold for all higher values of ϵ .

It is evident that all four of these histograms have entirely positive support, suggesting that the median dominance conditions hold even for $\epsilon = 0$. In the online appendix, we also show analogous histograms for the 25th and 75th quantiles with $\epsilon = 0.0$. There is overwhelming evidence that these histograms also have positive support and thus that the stronger SD condition is also likely to be true. As a second piece of

⁴Since we use only applicants who were summoned for interview, there is an additional level of selection that can further weaken the correlation between unobservables and observables. Our "test" therefore assesses the extent of correlation remaining after both levels of selection.

FIGURE 1.—EVIDENCE OF MEDIAN DOMINANCE



Histogram of differences in predicted median interview score across pairs of candidates where the first has scored higher than the second in terms of each of GCSE score, aptitude test, and essay. For example, the histogram in the top-left panel shows the estimated marginal distribution of the variable: $\text{Median}[\text{interview} | X_{\text{male}}, G = \text{male}] - \text{Median}[\text{interview} | X_{\text{female}}, G = \text{female}]$ across all paired realizations $(X_{\text{male}}, X_{\text{female}})$ satisfying $X_{\text{male}} \geq X_{\text{female}}$.

evidence, we calculate the correlation matrix among the various indicators of academic merit at the preadmission stage. These are reported in the online appendix, where it is evident that all correlations are strictly positive, which lends further support to assumption M/SD.

The evidence presented is of course suggestive rather than definitive. Indeed, if we had found a negative or no relation between the interview score and the observable test scores, our assumption M would be suspect. The point of figure 1 in the text and figures 8 and 9 and table 5 in the appendix is to show that this is not the case.

Our next assumption relates to the structure of the ϕ function.

Assumption CM (conditional monotonicity). (i) $\phi(x, g, z)$ is strictly increasing in z for every x and g ; (ii) if x_g and x_h satisfy $x_g \geq_\varepsilon x_h$, then $\phi(x_g, g, z) > \phi(x_h, h, z)$ for any z , and any $g \neq h$.

Part i of assumption CM is essentially definitional (regarding Z) in that higher values of the index of ability based on unobserved characteristics are associated with higher values of the perceived expected outcome. Part ii says that if a g -type applicant is better than an h -type applicant along a set of key observable characteristics *and* is at least equally good along the ability index, which is unobservable to us but observable to the decision makers, then the g -type applicant will be perceived to have a higher expected outcome by the decision maker. It is important for part ii that the g -type applicant is at least as good as the h -type applicant along

the index Z . For instance, suppose that admission tutors base their assessment on past exams whose scores X are observed by us and the quality of the reference letter Z , unobserved by us. Then a female candidate who has scored lower on every component of X than a male candidate but has a much better recommendation may or may not be perceived as having lower potential than the male candidate. But a female candidate who has an equally strong recommendation Z as a male candidate but has scored lower on every X than he did will likely be perceived to have lower academic potential in expectation. A sufficient but not necessary condition for CM (ii) to hold is that (a) $\phi(x, g, z) = \phi(x, h, z) \equiv \phi(x, z)$ for all x, z for any $g \neq h$ (i.e., conditional on the observable X and unobservable Z , the demographic characteristic G does not affect the outcome of interest), and, furthermore, (b) $\phi(x, z) \geq \phi(x', z)$ if $x \geq_\varepsilon x'$.

As a referee has pointed out, there is some evidence from California that women with lower SAT scores and high school GPA than men have performed systematically better in college examinations (Leonard & Jiang, 1999; Rothstein, 2004). This does seem somewhat unlikely in our application, given figures 1 and figure 3 below. Nonetheless, for the sake of robustness in our empirical application, we consider a variant of assumption CM where instead of the raw scores X_g and X_h , we use their standardized versions. That is, for group g , each performance measure X_g is taken not to be the raw score but as $X_g^{\text{con}} \equiv (\text{raw_score} - \mu_g) / \sigma_g$, where μ_g and σ_g are the mean and standard deviation of the raw score within group g . Accordingly, the condition $X_{\text{male}}^{\text{con}} \geq_\delta X_{\text{female}}^{\text{con}}$ refers to male-female pairs where the males have higher

relative scores than females: $\frac{X_{male} - \mu_{male}}{\sigma_{male}} \geq \frac{X_{female} - \mu_{female}}{\sigma_{female}} + \delta$. Then the contextual version of assumption CM (ii) is given by:

Assumption CM' (conditional contextual monotonicity).

- (i) $\phi(X, g, z) \equiv \phi(X^{con}, g, z)$, for all g, z ; the function $\phi(x^{con}, g, z)$ is strictly increasing in z for every x^{con} and g .⁵
- (ii) If x_g^{con} and x_h^{con} satisfy $x_g^{con} \geq_\delta x_h^{con}$, then $\phi(x_g^{con}, g, z) > \phi(x_h^{con}, h, z)$ for any z , and any $g \neq h$.

This assumption means that candidates whose performances are in the top echelons of their own sociodemographic group will be perceived to be academically stronger. It thus allows for biased performance measures—for example, that female applicants with lower raw scores on preentry evaluations may perform better on college exams, on average, and may therefore be favored by admission officers over men with higher initial scores. In our empirical work, we will report the results using both the raw and the standardized scores to compare pairs of applicants.

Choice of ε . A practical way of choosing ε is to draw histograms based on observables like figure 1 for a range of values of ε and then choose the smallest value for which the corresponding histograms have entirely positive support. In the application discussed below, we report results for $\varepsilon = 0.1$ and $\varepsilon = 0.25$ to ensure that there is no overlap in observable characteristics between the pairs of students compared. Indeed, from figure 1, it is obvious that any value of ε larger than 0 should be acceptable for this application. We also provide some robustness check by reporting results over a range of ε in figure 7.

B. Identification Analysis

We show how assumptions M/SD and CM can be used to identify the sign of threshold differences. To see this, denote the threshold used for type g and type h applicants by γ_g and γ_h , respectively. Under meritocratic admissions, one expects $\gamma_g = \gamma_h$. Define the function

$$p(x, g) := \Pr [D = 1 | X = x, G = g] \\ := \Pr [\phi(X, G, Z) > \gamma_g | X = x_g, G = g],$$

and the set $\mathcal{M}(g, h, \varepsilon)$ as

$$\mathcal{M}(g, h, \varepsilon) := \{ (x_g, x_h) \in \mathcal{X}_g \times \mathcal{X}_h : x_g \geq_\varepsilon x_h, p(x_g, g) \leq 0.5 < p(x_h, h) \}. \quad (2)$$

Note that the set $\mathcal{M}(g, h, \varepsilon)$ can be directly computed from the data because it depends only on observables.

⁵Part i of this assumption is identical to CM(i), since one can always rewrite $\phi(x, g, z) = \phi(\mu_g + \sigma_g x^{con}, g, z) \equiv \xi(x^{con}, g, z)$ with the monotonicity of $\phi(x, g, z)$ in x carrying over to monotonicity of $\xi(x^{con}, g, z)$ in x^{con} .

Now suppose that one finds that $\mathcal{M}(g, h, \varepsilon)$ is nonempty. Then, for any (x_g, x_h) in $\mathcal{M}(g, h, \varepsilon)$, since $p(x_g, g) = \Pr [\phi(x_g, g, Z) > \gamma_g | x_g, g] \leq 0.5$, it must be true that

$$\begin{aligned} \gamma_g &\geq \text{Median} [\phi(X, G, Z) | X = x_g, G = g] \\ &= \phi(x_g, g, \text{Median} [Z | x_g, g]), \text{ by assumption CM(i)} \\ &> \phi(x_h, h, \text{Median} [Z | x_g, g]), \text{ by CM(ii)} \\ &\geq \phi(x_h, h, \text{Median} [Z | x_h, h]), \text{ by assumption M} \\ &= \text{Median} [\phi(X, G, Z) | X = x_h, G = h], \text{ by CM(i)} \\ &\geq \gamma_h, \text{ since } 0.5 < p(x_h, h). \end{aligned}$$

Thus, the nonemptiness of the set $\mathcal{M}(g, h, \varepsilon)$ leads to the inequality $\gamma_g > \gamma_h$.

Under the stronger SD assumption, nonemptiness of the set

$$\mathcal{SD}(g, h, \varepsilon) := \{ (x_g, x_h) \in \mathcal{X}_g \times \mathcal{X}_h : x_g \geq_\varepsilon x_h, p(x_g, g) < p(x_h, h) \} \quad (3)$$

would analogously imply that $\gamma_g > \gamma_h$. This is because if $(x_g, x_h) \in \mathcal{SD}(g, h, \varepsilon)$, then because $1 - p(x_g, g) = \Pr [\phi(X, G, Z) < \gamma_g | X = x_g, G = g]$, we have that

$$\begin{aligned} \gamma_g &= Q^{1-p(x_g, g)} [\phi(X, G, Z) | X = x_g, G = g] \\ &= \phi(x_g, g, Q^{1-p(x_g, g)} [Z | x_g, g]), \text{ since } \phi(x_g, g, \cdot) \\ &\quad \text{is increasing} \\ &> \phi(x_g, g, Q^{1-p(x_h, h)} [Z | x_g, g]), \\ &\quad \text{since } p(x_g, g) < p(x_h, h) \\ &\geq \phi(x_g, g, Q^{1-p(x_h, h)} [Z | x_h, h]), \text{ by assumption SD} \\ &\quad \text{since } x_g \geq_\varepsilon x_h \\ &\geq \phi(x_h, h, Q^{1-p(x_h, h)} [Z | x_h, h]), \text{ by assumption CM(ii)} \\ &\quad \text{since } x_g \geq_\varepsilon x_h \\ &= Q^{1-p(x_h, h)} \{ \phi(x_h, h, Z) | x_h, h \}, \text{ since } \phi(x_h, h, \cdot) \\ &\quad \text{is increasing} \\ &\geq \gamma_h, \end{aligned}$$

since

$$1 - p(x_h, h) = \Pr \{ \phi(X, G, Z) < \gamma_h | X = x_h, G = h \}.$$

Intuitively speaking, here the identification-relevant information comes from those pairs of g -type and h -type applicants for whom the dominance condition $x_g \geq_\varepsilon x_h$ holds and yet the g -type's probability of being accepted is lower. Assumption M (or SD) guarantees that these g -type applicants are also better, in a stochastic sense, in terms of unobservables. Note that these identifying pairs include applicants who are close to each other (albeit at least ε standard deviations apart) in terms of observables and also those that are further apart. Also when $\gamma_g - \gamma_h > 0$, it must be the case that $\mathcal{SD}(h, g, \varepsilon)$ is empty. Therefore, if one finds

that $\mathcal{SD}(g, h, \varepsilon)$ is empty, then one may test if $\mathcal{SD}(h, g, \varepsilon)$ is nonempty. If so, then one can conclude that $\gamma_g < \gamma_h$.

Remark 5. The logical structure of our analysis is that if $\mathcal{S}^{SD}(g, h, \varepsilon)$ is nonempty, then we can conclude that $\gamma_g > \gamma_h$. But it is possible that although $\gamma_g > \gamma_h$, we find that $\mathcal{S}^{SD}(g, h, \varepsilon)$ is empty. This is a generic feature of any analysis based on partially identified parameters: they must be conclusive in fewer instances compared to when model parameters are point identified. In other words, the cost of allowing for unobservables is that we may lose the ability to detect very small but positive threshold differences, but when we detect a difference, we can be certain about its existence. Indeed, without our proposed methods and the underlying assumptions justifying them, one cannot in general detect any threshold difference however large they might be.

C. Alternative Identification Strategies

Our methodology may be contrasted with some alternative strategies proposed in the literature in noneducational contexts. For instance, in the context of health care, Chandra and Staiger (2009) attempt to identify differences in expected outcome thresholds for surgery by assuming an index restriction on the unobservable's distribution. This approach fails when the distribution of the unobservables differs across G , conditional on observables. Our analysis imposes no such restriction on the unobservables' distribution. In the health care context, Bhattacharya (2013) suggests an alternative approach to testing treatment bias using a combination of observational data and prior experimental findings from randomized controlled trials. Such experimental data are difficult to come by for college admissions. In law enforcement contexts, several researchers have relied on the assumption that target individuals react optimally to treatment protocols and devised methods to detect racial prejudice using this (Persico, 2009). However, these approaches rely on the specifics of the context and do not generalize to situations involving university admissions. For example, it is both difficult for university applicants to alter their potential academic outcomes in response to admission protocols and impractical for them to want to do this given the one-shot nature of admission exercise.

V. Estimation and Inference

Given the identification analysis, our next task is to develop a formal inference method for testing threshold differences. For this purpose, we will make the stronger assumption of SD rather than M. Indeed, these two assumptions have the same intuitive interpretation; the evidence for SD (see section VI and also part B of the appendix) is strong, and conducting statistical inference under it is slightly simpler.

The key task regarding inference, corresponding to assumptions SD and CM, is to test whether $\mathcal{SD}(g, h, \varepsilon)$, defined in equation (3),

$$\mathcal{SD}(g, h, \varepsilon) := \left\{ (x_g, x_h) \in \mathcal{X}_g \times \mathcal{X}_h : x_g \geq_\varepsilon x_h, p(x_g, g) < p(x_h, h) \right\}$$

is nonempty. Observe that the null hypothesis of an empty $\mathcal{SD}(g, h, \varepsilon)$ is equivalent to the hypothesis that $\alpha_0 \geq 0$, where

$$\alpha_0 := \inf_{(x_g, x_h) \in \mathcal{X}_g \times \mathcal{X}_h, x_g \geq_\varepsilon x_h} [p(x_g, g) - p(x_h, h)].$$

We will now outline how to test the emptiness of $\mathcal{SD}(g, h, \varepsilon)$, based on an inference method developed for "intersection bounds" by Chernozhukov, Lee, and Rosen (2013, hereafter, CLR). Although our identification method is nonparametric in the sense of not requiring functional form specifications, estimation and inference for the nonparametric case are complicated. Due to relatively small sample size, the two-sample nature of the problem, and the complicated construction of intersection bounds for nonparametric estimates (requiring subjective choice of various tuning parameters), we do not consider such methods here. Instead, we focus on the case where $p(\cdot, \cdot)$ is parametrically specified as a probit, that is,

$$p(x_g, g) = \Pr[D = 1 | (X, G) = (x_g, g)] = \Phi(x'_g \delta_{0,g});$$

$$\text{and } p(x_h, h) = \Phi(x'_h \delta_{0,h}),$$

where $(\delta_{0,g}, \delta_{0,h})$ are the probit coefficients and Φ is the CDF of the standard normal. Note that under our parametric specification, $\Phi(x'_g \delta_{0,g}) \leq \Phi(x'_h \delta_{0,h})$ is equivalent to $x'_g \delta_{0,g} \leq x'_h \delta_{0,h}$ and thus

$$\mathcal{SD}(g, h, \varepsilon) = \{x_g \geq_\varepsilon x_h, x'_g \delta_{0,g} \leq x'_h \delta_{0,h}\},$$

and thus emptiness of $\mathcal{SD}(g, h, \varepsilon)$ is equivalent to the hypothesis that $\theta_0 \geq 0$, where

$$\theta_0 := \inf_{(x_g, x_h) \in \mathcal{X}_g \times \mathcal{X}_h, x_g \geq_\varepsilon x_h} [x'_g \delta_{0,g} - x'_h \delta_{0,h}].$$

The quantity θ_0 is exactly of the form analyzed in CLR (2013). We construct a one-sided 95% confidence interval $\hat{C}_n(0.95) = (-\infty, \hat{\theta}_{n0}(0.95))$ for θ_0 by adapting the CLR method, as outlined in part C of the appendix, for each choice of g and h . If $\hat{\theta}_{n0}(0.95) < 0$, then we conclude that $\mathcal{SD}(g, h, \varepsilon)$ is nonempty.

VI. Empirical Analysis

We provide summary statistics for our sample in table 2. The left half of table 1 shows that male applicants have better aptitude test scores and interview averages. They perform slightly worse on average in their GCSE and A-levels. These differences are statistically significant at the 5% level. Note that there is no significant difference in offer rates between male and female candidates. The independent and state school applicants are quite similar in terms of most

TABLE 2.—MEANS BY GENDER AND SCHOOL TYPE

Variable	Female ($N = 241$) (1)	Male ($N = 394$) (2)	pvalue_diff (3)	State ($N = 355$) (4)	Independent ($N = 280$) (5)	pvalue_diff (6)
gcsescore	3.79	3.72	0	3.67	3.85	0
alevelscore	119.73	119.59	0.01	119.60	119.73	0.02
aptitude test	62.02	65.09	0	63.16	64.85	0.0015
essay	61.77	63.38	0	62.98	64.42	0.5
interview	63.74	64.69	0.04	64.24	64.43	0.65
prelim_avg	61.02	62.33	0.04	61.83	61.83	0.03
offer	0.33	0.37	0.14	0.34	0.35	0.24
accept	0.33	0.37	0.5	0.34	0.35	0.46

The data pertain to two cohorts of applicants. The variable names are explained in table 1. Columns 3 and 6 record the p -value corresponding to a test of equal means against a one-sided alternative. Differences in unconditional offer rates across school types (highlighted) are seen to be statistically indistinguishable from 0 at 5%.

TABLE 3.—PROBIT REGRESSION OF ADMISSION

Variable	Coefficient	Coefficient/ SE	Marginal Effect	Marginal Effect/SE
gcsescore	0.188	0.76	0.055	0.75
interview	0.225	10.43	0.066	11.72
aptitude test	0.087	6.99	0.026	6.76
essay	0.007	0.59	0.002	0.59
male	-0.210	-1.33	0.062	-1.31
independent	-0.129	-0.84	0.037	-0.84

Probit regression of eventual admission for all U.K.-based applicants, together with two-sided p -value. The highlighted fields show insignificant effect of gender and school background on admission probabilities, controlling for aptitude test scores. Data pertain to two cohorts of U.K.-based applicants. Marginal effects are calculated at mean values of covariates and for moving from 0 to 1 for male and independent. Gender and school type remain insignificant (highlighted in italics) even after controlling for past test-scores.

characteristics except for a slightly higher GCSE for the former.

In table 3, we report the results of a probit regression of receiving an offer across all applicants. Table 3 strengthens the findings from table 2 by showing that even after controlling for covariates, gender and school type do not affect the average admission success rate among applicants. The value of McFadden's pseudo- R^2 for the probit model is about 50%, and the corresponding R^2 for a linear probability model (not reported here) is about 45%, about ten times higher than the goodness-of-fit measures typically reported by applied researchers working with cross-sectional data. This suggests that the commonly observed covariates explain a very large fraction of admission outcomes. Moreover, table 3 also shows that the aptitude test and interview scores have the largest impact upon receiving an offer for the applicant population (in terms of the t -statistics).

A. Results

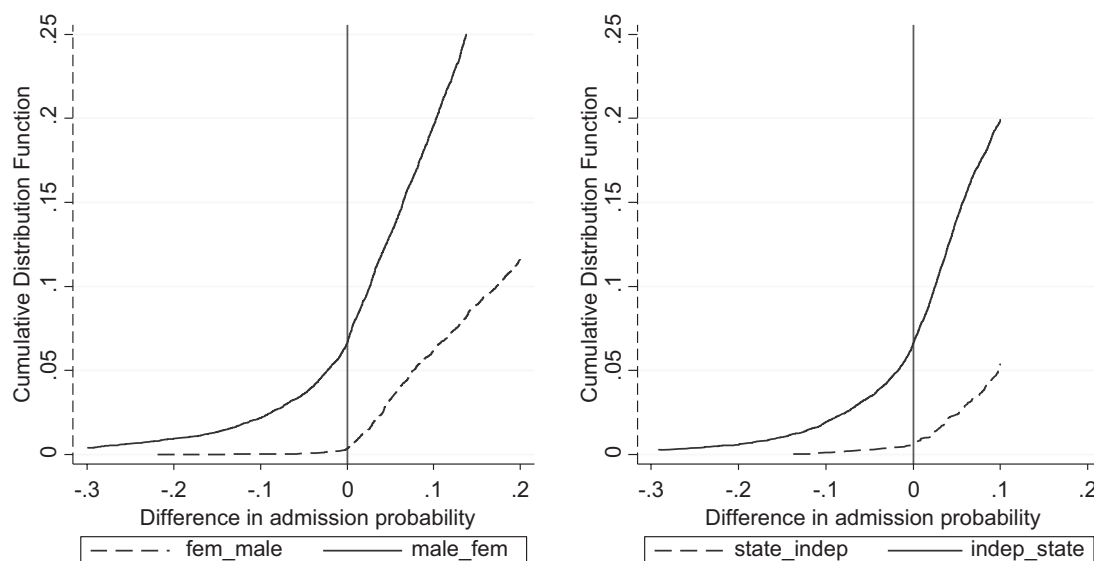
We first conducted a simulation exercise, reported in the online appendix, part B.3, using these data, to check if our methods work well in a setting where we “know” the true admission criteria. In that exercise, we find that medium-sized differences in admission thresholds are picked up by our method and very small differences are not, which increases our confidence that the methods work well in practice. Now we turn to the real application where we use the gcsescore, aptitude test score, and the interview score as the covariates X for defining dominance. That is, if a g -type

candidate has scored ε standard deviations higher on each of these three key assessment scores than an h -type candidate, then the conditional distribution (or median) of the unobservable component of assessment for the former is assumed to dominate that for the latter for all g and h , as per assumption M or SD.

In accordance with the discussion in section V, the first step is to examine the emptiness of $\mathcal{SD}(g, h, \varepsilon)$ using data on only X and D . We first investigate this graphically in figure 2 by plotting the marginal CDF of the difference in admission probabilities $p(X_g, g) - p(X_h, h)$ for pairs of (X_g, X_h) satisfying $X_g \geq_\varepsilon X_h$ for $\varepsilon = 0.1$ for various combinations of g and h . The predicted probabilities $p(\cdot, \cdot)$ are calculated separately for each group g via standard probit using gcsescore, aptitude test score, the examination essay score, and the interview score as regressors. Since we concluded dominance with $\varepsilon = 0.0$, with Z being the interview score, we chose a slightly higher (i.e., more conservative) value of $\varepsilon = 0.1$ to investigate the emptiness of $\mathcal{SD}(g, h, \varepsilon)$. When the event $\{X_g \geq_\varepsilon X_h\}$ happens with positive probability, an empty $\mathcal{SD}(g, h, \varepsilon)$ is equivalent to $\Pr[X_g \geq_\varepsilon X_h, p(X_g, g) < p(X_h, h)] = 0$, where the probability is taken with respect to the distributions of X_g and X_h . Therefore, a positive mass at and below 0 for these CDFs indicates that $\mathcal{SD}(g, h, \varepsilon)$ is nonempty. In the left panel, when $g = \text{male}$, $h = \text{female}$, the CDF is represented by the solid curve, labeled male_fem, and when $g = \text{female}$ and $h = \text{male}$, it is the dashed curve, labeled fem_male. A positive height at 0 indicates that applicants with higher observables in the first group have lower admission probabilities than the second.

Clearly the first curve has significant mass below 0 and the dashed curve has almost no mass below 0, suggesting a positive probability that $p(X_{\text{male}}, \text{male}) < p(X_{\text{female}}, \text{female})$ although $X_{\text{male}} \geq_\varepsilon X_{\text{female}}$. This evidence is still present in the right panel, with independent and state schools replacing male and female, respectively, but to a slightly lesser extent, suggesting that γ_{indep} may be only slightly larger than γ_{state} . To perform the test formally, in table 4, we report $\hat{\theta}_{0n}(0.95)$, the upper limit of a one-sided confidence interval, calculated using the method of CLR, as explained in section V. We report results for $\varepsilon = 0.1$ (recall that we concluded dominance even with $\varepsilon = 0$; see figure 2). A negative

FIGURE 2.—GRAPHICAL EVIDENCE OF DIFFERENT ADMISSION THRESHOLDS



The graphs plot the marginal CDF of the difference in admission probabilities $p(Xg, g) - p(Xh, h)$ for pairs of (Xg, Xh) satisfying $Xg > \epsilon Xh$ for $\epsilon = 0.1$ for various combinations of g and h . A positive height at 0 indicates that applicants with higher observables in the first group (g) have lower admission probabilities than those with lower observables in the second group (h). The solid curve on the left panel shows, for example, that a subgroup of men with higher observables have a lower admission probability than a subgroup of women with lower observables.

TABLE 4.—TESTING UNEQUAL THRESHOLDS

Difference	$\epsilon = 0.1$	$\epsilon = 0.25$	Quadratics in Preadmission Scores, $\epsilon = 0.1$	Standardized Scores, $\delta = 1.5$
$g = \text{male}, h = \text{female}$	-1.73	-2.02	-3.49	-2.01
$g = \text{female}, h = \text{male}$	0.57	0.67	0.684	0.43
$g = \text{indep}, h = \text{state}$	-1.29	-0.58	-2.75	0.012
$g = \text{state}, h = \text{indep}$	0.92	0.04	0.635	1.87
$g = \text{state_male}, h = \text{state_female}$	-1.36	-1.01	-6.85	-1.19
$g = \text{indep_male}, h = \text{indep_female}$	-1.11	-3.39	-2.7	-3.56

This table reports the upper limit of the one-sided 95% confidence interval for testing whether group g is facing a higher admission threshold than group h , with a negative upper limit indicating that it is. The first two columns (with $\epsilon = 0.1$ and $\epsilon = 0.25$) correspond to evaluating difference in admission probability (as a function of g score, aptitude test score, essay score, and interview score) between a g -type and an h -type applicant where the former has scored ϵ standard deviations higher on each of the raw preentry performance measures, and the final column corresponds to the case where the former has scored 1.5 points or higher on standardized Z-score versions of them, as explained in the text in section 6.2a. The penultimate column shows the results when quadratics and second-order interactions between all preadmission performance measures are used as additional controls, beyond the linear versions of them, to predict admission probabilities, as a robustness check.

upper limit indicates that the set $SD(g, h, \epsilon)$ is nonempty; consequently, we reject the null of $\gamma_g \leq \gamma_h$ in favor of $\gamma_g > \gamma_h$. It is evident from the first four rows of table 4 that we reject emptiness for $g = \text{male}, h = \text{female}$ and for $g = \text{indep}, h = \text{state}$ but not in the other cases. This suggests that males and private school applicants face higher admission thresholds. The exact upper limits of confidence intervals reported vary slightly across functional specifications—for example, whether higher-order terms and interactions in the test scores are or are not used to estimate $p(\cdot, \cdot)$ —but two empirical findings are robust across all specifications: (a) the gender gap is large, persistent, and statistically significant in every case,⁶ and (b) the independent–state school

difference is less persistent across specifications but is always negative. Given the evidence of a large gender gap, we investigated it further by breaking the data up by school type. Results reported in the last two rows of table 4 show that the gender gap is large within both state and private school categories, indicating that male applicants are held to a higher standard for applicants from both state and private school backgrounds.

It would be natural to conjecture that the threshold differences arise primarily from the implicit or explicit practice of affirmative action on the overweighting of outcomes for historically disadvantaged groups. A second possibility is that in face of political or media pressure, admission tutors try to equate an application success rate for, say, men with one for women, which is also consistent with our empirical findings (see tables 2 and 3). This would make the effective male threshold higher if, say, the conditional male outcome distribution has a thicker right tail. A third possibility is that female applicants have a lower admission threshold in order to encourage more women to apply in the future. Note from

⁶As noted by a referee, this finding is somewhat curious, given that girls routinely outperform boys in the majority of high school and college tests across the world, including the PISA assessments (Goldin, Katz, & Kuziemko, 2006, and Niederle & Vesterlund, 2010). Indeed in our data, the performance of the average (as opposed to marginal) female admit is also lower than that of the average male admit, although this has nothing to do with admission thresholds and fair admission per se.

TABLE 5.—REGRESSION OF FIRST-YEAR PERFORMANCE ON OBSERVABLE COVARIATES

	Coefficient	SE	<i>t</i> -value
gcsescore	3.33	1.77	1.88
aptitude test	0.19	0.04	4.31
essay	−0.004	0.047	−0.08
interview	0.06	0.03	1.78
male	1.14	<i>0.69</i>	<i>1.66</i>
indep	0.41	<i>0.68</i>	<i>0.75</i>

Regression of admitted candidates' performance on first-year examinations on preadmission test scores, interview score, gender, and school type. Italicized fields show significant positive impact of being male but insignificant effect of being from a private school on subsequent performance, conditional on admission.

table 2 that the number of female applications is nearly half the number of male ones. Regardless of what the underlying determinants of the tutors' behavior are, we can conclude from our analysis that the admission practice under study deviates from the outcome-oriented benchmark and makes male and independent school applicants face significantly higher admission thresholds.

In order to gain some further insight into how the threshold discrepancies arise, we plot the empirical CDFs of predicted academic performance based on the observable characteristics. This is done by regressing first-year and then final-year examination scores in university on gcsescore, aptitude test and essay score, interview grades, and gender or school type for enrolled students. The regression output appears in table 5. The estimated CDFs of predicted performance by gender and school type are plotted in figure 3B.

It is clear that in both graphs, the male distribution first-order stochastically dominates the female distribution. This means that if admissions were determined solely by predicted performance based on observables (i.e., there is no unobserved heterogeneity), any common acceptance rate across gender will result in a higher predicted outcome for the marginal accepted male than the marginal accepted female. The dominance is less pronounced in the case of school type, since female independent school candidates appear to face a lower threshold than male state school candidates. Our results in table 4 imply that allowing for unobserved heterogeneity does not change this scenario substantively and suggests that equating the application success rates (see table 2) leads to the use of higher admission thresholds for male and, to a lesser extent, private school candidates. Indeed, if admission officers believe that eventual exam performance is not the relevant measure of merit, then one needs to repeat the analysis with whichever performance measure "meritocracy" is defined with respect to. Taking the attainment of at least a 2.1 (a high-second-class mark of 64%), a minimum requirement for entry into most postgraduate programs, the relevant outcome produces a very similar result, presented in figure 4.

At this point, it is worth considering whether our findings could be consistent with two other alternative explanations, as follows.

G-blind admissions. The first possibility is where admission tutors ignore G completely in forming their assessment and use a common admission cutoff across G , thereby generating insignificant effects of gender and school type on admission probabilities, both unconditionally (see table 2) and conditionally on past test scores (see table 3). Such behavior could arise from either an institutional norm banning any conditioning on demographic characteristics or from the tutors' belief that such characteristics have no explanatory power beyond the preadmission test scores. Therefore, the question is whether by including G in our analysis, we are detecting threshold differences that are not "intentional." Even if that were the case, we would argue that in order for admissions to be meritocratic, admission tutors should take G into account. For example, suppose G denotes a school type—state school students are more able than independent school students with the same test score and therefore perform better in postadmission exams. If tutors ignore G , then an independent and a state school student with identical preadmission test scores will have an equal probability of admission, even though the state school student is more meritorious, which would contradict the notion of meritocratic admissions.

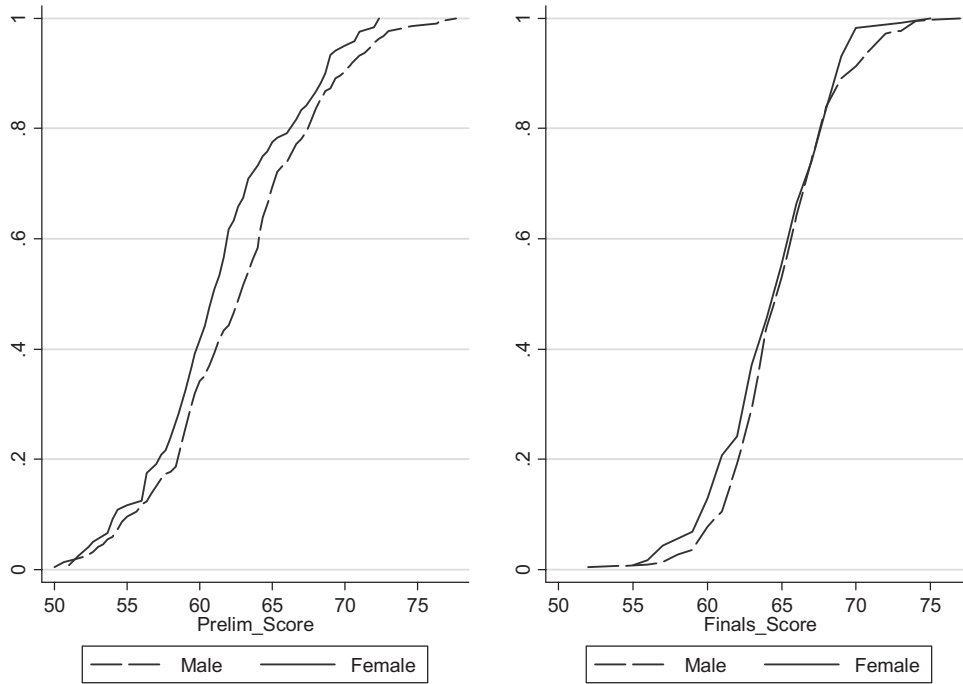
Biased interviews score. A second issue concerns the use of interview scores in calculating the lower bounds. Suppose that tutors are biased in favor of type g applicants and award them higher interview scores (relative to true performance) than type h . But as we saw in figure 1, the interview score does appear to satisfy assumption M (with $\varepsilon = 0$), which would be unlikely if one type of candidate was systematically awarded higher interview scores relative to their performance in the other more "objective" tests. For example, for $g = \text{male}$ and $h = \text{female}$, if men are awarded systematically higher interview scores, then we would expect to see a significant mass in the negative orthant of the top-right histogram in figure 1, which is clearly not the case.

B. Some Robustness Checks

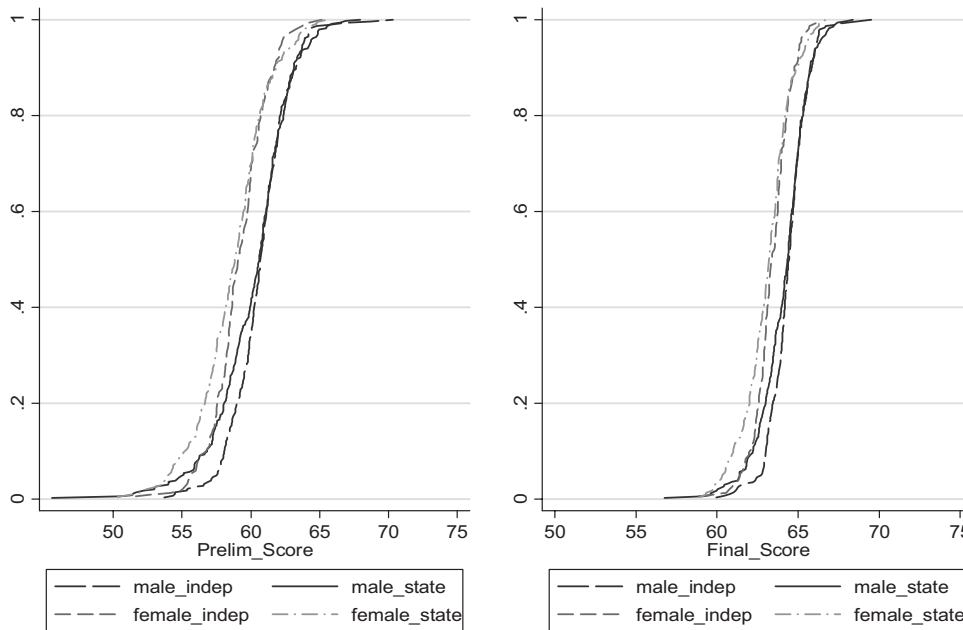
Biased test scores. One feature of our approach is that we are taking the preadmission test scores as true indicators of academic merit. However, students from privileged backgrounds might perform well in these tests simply on account of having been coached extensively. It is not possible to conduct any analysis of meritocracy if no previous measure of achievement can be taken to be accurate. Postadmission performance is not observed for nonadmitted candidates and thus cannot be incorporated in the analysis without strong assumptions. Therefore, it is important to examine whether our substantive conclusions are affected if we use contextualized (i.e., standardized) scores within each demographic group as an alternative measure of merit (figure 5). Accordingly, we repeat the analysis by replacing each test score by its standardized version and invoking assumption CM' .

FIGURE 3.—CDF OF ACTUAL AND PREDICTED FIRST-YEAR (PRELIMINARY) AND THIRD-YEAR (FINAL) PERFORMANCE

A. CDFs of First-Year (left) and Final-Year (right) Performance in College for Admitted Candidates, by Gender



B. CDFs of Predicted First-Year (Preliminary) and Third-Year (Final) Performance Based on Observables, by Gender and School Type



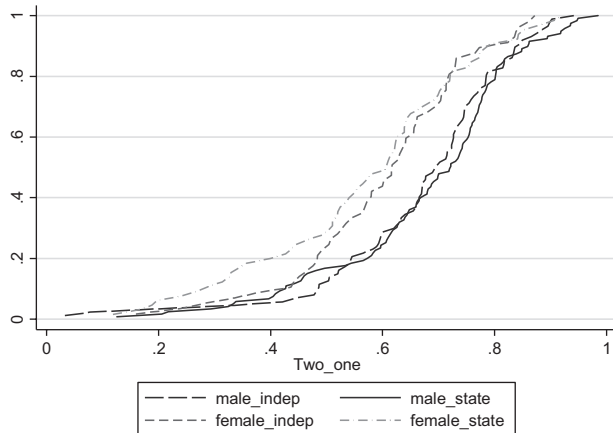
(A) The male CDFs lie almost entirely to the right of the female CDFs, with dominance more pronounced for preliminary performances. (B) CDFs for admitted candidates based on observable GCSE score, interview performance, and aptitude test scores. The male CDFs are seen to lie almost entirely to the right of the female CDFs, implying that a common admission rate would imply that marginal male entrants will have significantly higher expected score on first- and final-year exams.

Recall the condition $X_{male\delta} \geq_{\delta} X_{female}$, which refers to male-female pairs where the males have higher relative scores than females. Then we can conclude that group g faces a higher threshold than group h if $\theta_0 < 0$, where

$$\theta_0 \equiv \inf_{(x_g, x_h) \in \mathcal{X}_g \times \mathcal{X}_h, x_g \geq_{\delta} x_h} [x'_g \delta_{0,g} - x'_h \delta_{0,h}].$$

The results from this exercise are shown in table 4, in the last column, “Standardized Scores,” corresponding to $\delta = 1.25$ (the smallest δ for which histograms analogous to those in figure 1 have positive support). As before, a negative upper limit of the CLR confidence interval indicates that group g faces a higher threshold than group h , since some group g

FIGURE 4.—PREDICTED PROBABILITY OF ATTAINING A SCORE OF 2.1



The graph plots the CDF of the predicted probability of getting at least a high second-class score (64%) on first-year exams, based on aptitude test score, interview score, and GCSE score. The horizontal axis marks the probability of getting at least a 2.1, and the vertical axis is the admission probability. A common success rate for entry would imply a lower threshold for female and state school candidates, but with male state school candidates facing a higher threshold than female independent school candidates. For instance, a 30% success rate across school type and gender would imply that about 63% of female candidates from state schools and about 75% of male private school candidates would get at least a 2.1 degree in expectation. This figure is a robustness check on figure 3.

members with high relative test scores have a lower probability of admission than some group h members with lower relative test scores. As is apparent from table 4, in the last column, it still remains the case that male applicants face a higher admission threshold than female candidates. However, the test for a threshold difference between independent and private school students now becomes inconclusive. This confirms the previous substantive finding that threshold differences by school type are insignificant but the gender differences are pronounced.

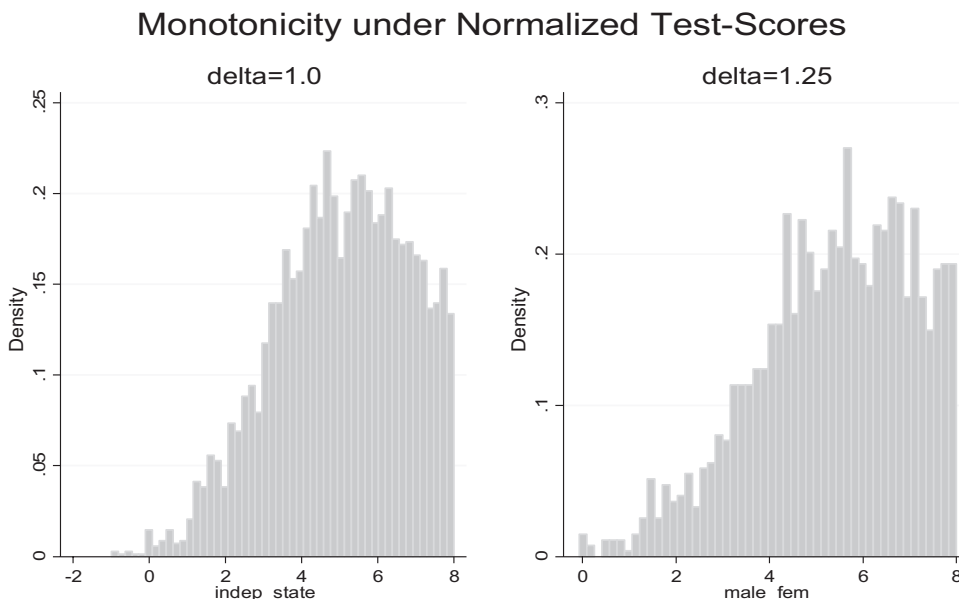
First-stage selection. In principle, we can repeat our analysis to test meritocracy in the first-stage selection process as well. However, the first-stage selection in our empirical context is based entirely on the ranking in the aptitude test scores; there is effectively no selection on unobservables at this stage. In particular, all applicants are classified into bands by their overall aptitude test score. Then private school students in approximately the top half and state school students in the top two-thirds are interviewed. Figure 6 presents suggestive evidence regarding the first-stage selection of candidates. The left panel plots the CDF of aptitude test scores for those making it to the interview stage. The right graph plots the CDF of predicted interview scores based on the aptitude test score (analogous to figure 3 for the second stage of selection). A common success rate for entry to the interview stage would imply a lower threshold for female and state school candidates, but with male state school candidates facing a higher threshold than female independent school candidates. Thus, in fact, one sees a very similar overall picture as in the second-stage selection (see figure 3).

Choice of ϵ . Finally, in figure 7, we plot the upper limits of the CLR confidence intervals across a range of ϵ for both the overall gender gap, as well as the gender gap within each school type. The persistence of the negative upper limits in figure 7 reinforces the conclusion that female candidates face lower thresholds than males both on average and within each type of school background.

C. Caveats

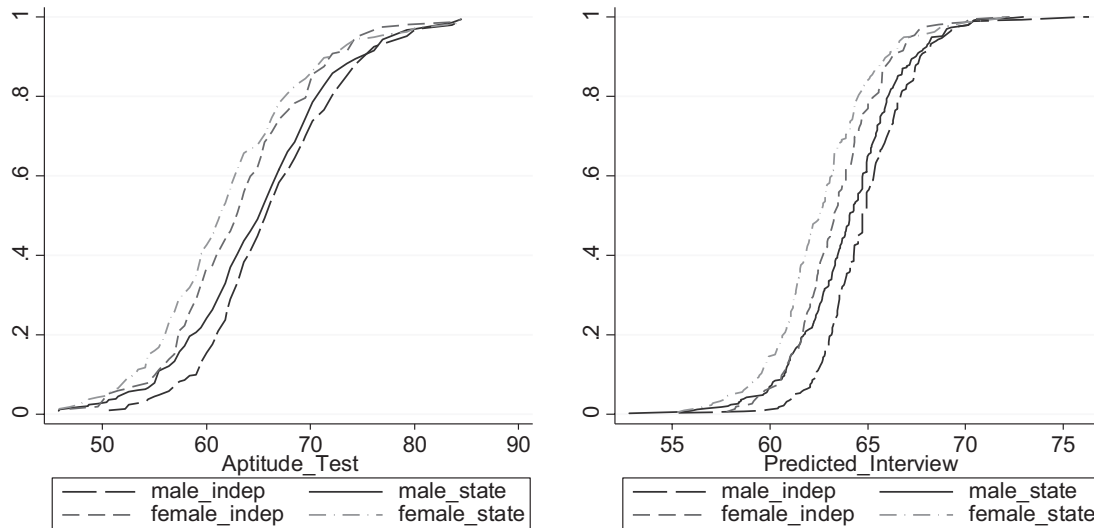
Several caveats apply to our methods and data. The first is that we ignore peer effects at both the individual and the

FIGURE 5.—TESTING MONOTONICITY OF MEDIAN INTERVIEW SCORE IN CONTEXTUALIZED TEST SCORES



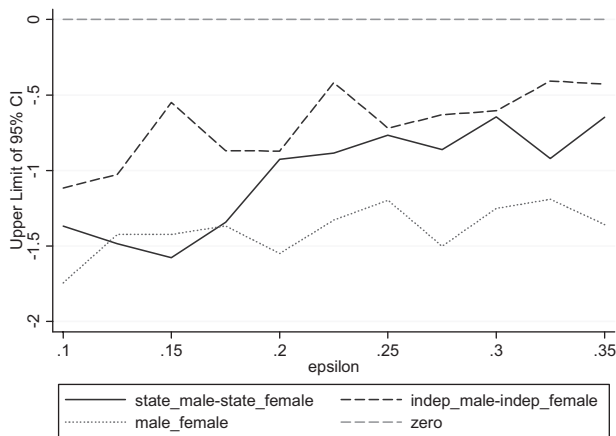
The graphs plot histograms of the difference in admission probabilities $p(X_{male}, male) - p(X_{female}, female)$ for pairs of (X_{male}, X_{female}) satisfying $X_{male} > X_{female} + \delta$, for $\delta = 1.0$ and $\delta = 1.25$, where X_{male}, X_{female} are the standardized test scores observed prior to admission. The smallest δ for which these histograms have positive support is $\delta = 1.25$. We use this value of δ to do our robustness checks.

FIGURE 6.—FIRST-STAGE SELECTION



The graphs present suggestive evidence regarding first-stage selection of candidates. (Left) The CDF of the raw aptitude test scores for those making it to the interview stage. (Right) The CDF of predicted interview scores based on aptitude test score and GCSE score, analogous to figure 3, which pertains to the second stage of selection. A common success rate across gender and school type for entry to the interview stage would imply a lower threshold for female and state school candidates, but with male state school candidates facing a higher threshold than female independent school candidates.

FIGURE 7.—EFFECT OF ϵ ON GENDER GAP IN ADMISSION THRESHOLDS



We examine how the overall male-female gap in thresholds differs by school type and how the results are affected by one's choice of ϵ . We plot upper limits of 95% CLR confidence intervals, with a negative upper limit implying that the first group faces a higher threshold than the second. These limits are plotted across a range of ϵ .

institutional levels. For example, it is possible that an applicant is admitted (or rejected) because he or she is deemed to have the potential to create positive (negative) externalities on his or her peers' performance. But it seems unlikely to us that admission tutors can be confident enough in predicting peer effects for this consideration to play a significant role in admissions. Nonetheless, there remains a possibility that some students are admitted simply because they come from demographic groups that "fit better" with the institution, although their test scores might be lower. Indeed, if future academic performance is an index of that fit, then figures 3 and 5 do not support these possibilities. But of course, the fit may be judged with respect to other indices, and thus this caveat remains.

The second caveat pertains to the data we use. In reality, different applicants in our context are assessed by different tutors, each assessing a set of applications. But there is a significant reallocation of files across tutors to ensure that meritorious candidates are not excluded simply because the tutor assessing their files happened to have assessed a disproportionately large number of strong applicants. However, the reallocation of files need not be perfectly managed. Therefore, our test should be viewed as one of meritocratic admission at the level of the university as a whole, and deviations from it should be interpreted as having arisen from a variety of possible sources including explicit affirmative action, inefficient reallocation of files, and systematically incorrect beliefs of tutors.

A third possibility is that in other contexts (notably in the United States), it has been found that women perform better on college exams than men with the same preadmission test scores. If that were true, admission officers may admit female applicants who have scored relatively lower on preadmission assessments. This is unlikely to be the case in our application; indeed, figure 3A shows that postentry college performance of males' first-order stochastically dominates that of females, which is inconsistent with the explanation of superior female performance. Moreover, figure 3B shows that predicted college performance on the basis of observables is also stochastically higher for men, which provides further evidence against that explanation. However, when applying our methods to other contexts, it would be advisable to draw graphs analogous to figures 3A and 3B as a preliminary check.

VII. Summary and Conclusion

This paper has proposed an empirical approach to testing, on the basis of microdata, whether an existing admission

protocol is meritocratic, when a researcher observes some but not all applicant-specific information observed by admission tutors. The approach works by defining meritocratic admissions through a threshold-crossing model and then using admission data to learn the sign of the difference in admission thresholds for different demographic groups. These quantities are robust to the unobserved characteristics problem, under an intuitive assumption about the ranking of applicants by unobservable attributes. Applying our methods to admissions data for a selective U.K. university, we find that admission thresholds that male applicants face are significantly higher than for females, while those for private school applicants are possibly slightly higher than for state school applicants. In contrast, average admission rates are nearly identical across gender and across school type. These conclusions hold up to a large variety of robustness checks, as described in section VIC. Beyond the application to college admissions, our methods are potentially useful for testing the fairness of other binary decisions such as loan approval or surgery referrals, where allegations of bias are common.

REFERENCES

- Angrist, J., and W. Evans, "Children and Their Parents' Labor Supply: Evidence from Endogenous Variation in Family Size," *American Economic Review* 88 (1988), 450–477.
- Arcidiacono, Peter, "Affirmative Action in Higher Education: How Do Admission and Financial Aid Rules Affect Future Earnings?" *Econometrica* 73 (2005), 1477–1524.
- Arcidiacono, Peter, and Mike Lovenheim, "Affirmative Action and the Quality-Fit Trade-Off," NBER working paper 20962 (2015).
- Becker, Gary, *The Economics of Discrimination* (Chicago: University of Chicago Press, 1957).
- Bertrand, Marianne, Remma Hanna, and Sendhil Mullainathan, "Affirmative Action in Education: Evidence from Engineering College Admissions in India," *Journal of Public Economics* 94:1–2 (2010), 16–29.
- Bhattacharya, Debopam, "Evaluating Treatment Protocols Using Data Combination," *Journal of Econometrics* 173 (2013), 160–174.
- Card, David, and Alan Krueger, "Would the Elimination of Affirmative Action Affect Highly Qualified Minority Applicants? Evidence from California And Texas?" *Industrial and Labor Relations Review* 58 (2005), 416–434.
- Chandra, Amitabh, and Doug Staiger, "Identifying Provider Prejudice in Medical Care," Harvard University and Dartmouth College mimeograph (2009).
- Chernozhukov, Victor, Simon Lee, and Adam Rosen, "Intersection Bounds: Estimation and Inference," *Econometrica* 81 (2013), 667–737.
- Espenshade, Thomas, C. Y. Chung, and J. L. Walling, "Admission Preferences for Minority Students, Athletes, and Legacies at Elite Universities," *Social Science Quarterly* 85 (2004), 1422–1446.
- Fryer Ronald, and Glenn Loury, "Affirmative Action and Its Mythology," *Journal of Economic Perspectives* 19 (2005), 147–162.
- Goldin, Claudia, Larry Katz, and Ilyana Kuziemko, "The Homecoming of American College Women: The Reversal of the College Gender Gap," *Journal of Economic Perspectives* 20 (2006), 133–156.
- Graddy, Kathleen, and Margaret Stevens, "The Impact of School Inputs on Student Performance: An Empirical Study of Private Schools in the United Kingdom," *Industrial and Labor Relations Review* 58 (2005), 435–451.
- Greenstone, Michael, and Tom Gayer, "Quasi-Experimental and Experimental Approaches to Environmental Economics," *Journal of Environmental Economics and Management* 57 (2001), 21–44.
- Heckman, James, "Detecting Discrimination," *Journal of Economic Perspectives* 12 (1998), 101–116.
- Herrnstein, Richard, and Charles Murray, *The Bell Curve* (New York: Free Press, 1994).
- Hinrichs, Peter, "The Effects of Affirmative Action Bans on College Enrollment, Educational Attainment, and the Demographic Composition of Universities," this REVIEW 94 (2012), 712–722.
- Hoxby, Caroline, "The Changing Selectivity of American Colleges," *Journal of Economic Perspectives* 23:4 (2009), 95–118.
- Hurwitz, Michael, "The Impact of Legacy Status on Undergraduate Admissions at Elite Colleges and Universities," *Economics of Education Review* 30 (2011), 480–492.
- Kane, Thomas, "Racial and Ethnic Preference in College Admissions," in Christopher Jencks and Meredith Phillips, eds., *The Black-White Test Score Gap* (Washington, DC: Brookings Institution, 1998).
- Keith, Simon, R. M. Bell, A. G. Swanson, and A. P. Williams, "Effects of Affirmative Action in Medical Schools: A Study of the Class of 1975," *New England Journal of Medicine* 313 (1985), 1519–1525.
- Leonard, David, and J. Jiang, "Gender Bias and the College Predictions of the SATs: A Cry of Despair," *Research in Higher Education* 40 (1999), 375–407.
- Manski, Charles, "Identification of Binary Response Models," *Journal of the American Statistical Association* 83 (1988), 729–738.
- Niederle, Muriel, and Lise Vesterlund, "Explaining the Gender Gap in Math Test Scores: The Role of Competition," *Journal of Economic Perspectives* 24 (2010), 129–144.
- Persico, Nicola, "Racial Profiling? Detecting Bias Using Statistical Evidence," *Annual Review of Economics* 1 (2009), 229–254.
- Rothstein, Jesse, "College Performance Predictions and the SAT," *Journal of Econometrics* 121 (2004), 297–317.
- Sander, Richard, "A Systemic Analysis of Affirmative Action in American Law Schools," *Stanford Law Review* 57 (2004), 367–483.
- Waters, S., S. Cooper, N. Kuncel, P. R. Sackett, and J. Arneson, "Socio-economic Status and the Relationship between the SAT and Freshman GPA: An Analysis of Data from 41 Colleges and Universities" (2009), research.collegeboard.org.
- Zimdars, Anna, and Anthony Heath, "Elite Higher Education Admissions in the Arts and Sciences: Is Cultural Capital the Key?" *Sociology* 4 (2009), 648–666.

This article has been cited by:

1. Thomas McGregor, Brock Smith, Samuel Wills. 2019. Measuring inequality. *Oxford Review of Economic Policy* 35:3, 368-395. [[Crossref](#)]
2. Cristina Borra, Almudena Sevilla. 2019. COMPETITION FOR UNIVERSITY PLACES AND PARENTAL TIME INVESTMENTS: EVIDENCE FROM THE UNITED KINGDOM. *Economic Inquiry* 57:3, 1460-1479. [[Crossref](#)]
3. Daniel Lee. 2018. Does Implicit Bias Predict Dictator Giving?. *Games* 9:4, 73. [[Crossref](#)]