Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

Are University Admissions Academically Fair?

Debopam Bhattacharya* Shin Kanaya

University of Cambridge University of Aarhus

Margaret Stevens

University of Oxford

April 9, 2016.

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 that are better-qualified on observables would, on average, appear academically stronger to admission-tutors based on unobservables. This assumption reveals the sign of differences in admission-standards across demographic groups which are robust to omitted characteristics. Applying our methods to admissions-data from a British university, we find higher admission standards for males and slightly higher ones for private-school applicants, despite equal admission success-probability across gender and school-background.

Keywords: University admissions, affirmative action, economic efficiency, marginal admit, unobserved heterogeneity, threshold-crossing model, conditional stochastic dominance, partial identification.

JEL Codes: I23, J14, C35

*Address for correspondence: Debopam Bhattacharya, Department of Economics, University of Cambridge, United Kingdom. Email: debobhatta@gmail.com

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

1 Introduction

Admission practices at selective universities generate considerable public interest and political controversy, due to their close connection with inter-generational mobility and social discrimination. For example, in the UK 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. On the other hand, background-based admission quotas such as caste-based reservation in India and race-based affirmative action in the US have generated intense public controversy. Despite significant public interest in these issues, rigorous methods for modelling and testing "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 (c.f., Espenshade et al, 2004, Zimdars et al, 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, c.f., Heckman, 1998. For instance, if female candidates ceteris paribus perform better on interviews, and interview scores are unobserved by a researcher, then equal admission rate of observationally similar male and female candidates implies bias against female applicants. Indeed, in the empirical context investigated in the present paper, we find that socioeconomic backgrounds do not have statistically significant effects on admission rates, once we control for pre-admission 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 is less significant.

Beyond their obvious legal and political significance, such findings also have important

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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 post-admission assessments, e.g., college exams, as the marginal admit 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 admit to perform better than the marginal independent school admit. 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-tutors 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 spotting who the marginal admits are and, therefore, assessing whether the decision-maker acted in an academically unbiased way. Problems of this type been recognized by previous researchers in the context of detecting taste-based discrimination in hiring (c.f. Heckman, 1998). In the present paper, we devise a test for meritocratic admissions – based on the *differences* in admission-thresholds faced by different demographic groups – which, 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, i.e., applicant characteristics observed by admission-tutors but unobserved by the researcher. In our model, academic fairness corresponds to using identical thresholds of expected future performance across applicants from different demographic groups. Our key assumption – for which we will

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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 tutor, 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 using this assumption in conjunction with pre and post enrolment 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 UK 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.

Literature: A large volume of research exists in educational statistics on the analysis of admissions to selective colleges, focusing mainly on the United States (c.f. 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 rate 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 US law-school admissions. This is in contrast to many other studies – both academic and policy-oriented – which compare either average pre-admission test-scores (c.f. Herrnstein and Murray, 1994) or average post-admission performance across *all* (as opposed to marginal) admitted students from different

[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.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

socioeconomic groups (c.f. Keith et al., 1985, Sackett et al., 2009, Kane and William, 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 US college admission; Card and Krueger (2005) describe the reduced-form impact of eliminating affirmative action on minority students' application behavior in California; Hinrichs (2012) examines effects of banning preferential admission policies on enrolment patterns of both minority and non-minority students. Arcidiacono and Lovenheim (2015) provide a review of the empirical evidence on the effect of affirmative action on student-college mismatch. The present paper, though substantively related to the above works, has a different goal , viz., here we construct a formal econometric model where affirmative-action (or taste-based discrimination) and meritocracy have different empirical implications, and use it in conjunction with admissions-related micro-data to detect deviations from meritocracy. To our knowledge, the only other work in this literature which focuses on marginal admits is Bertrand et al (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 set-up 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 vis-a-vis 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 the present paper's methodology allows for the possibility that

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

approvers may observe a richer set of applicant characteristics than the researcher. Applying our methods one can then test whether the observed data are consistent with meritocratic approval processes, e.g., 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 2 sets up a simple theoretical model, followed by the corresponding empirical model of meritocratic admissions; Section 3 describes the data. Section 4 states the assumptions, provides empirical evidence in support of the key identifying assumption, and lays out the identification analysis. Section 5 discusses inference. Section 6 reports the empirical findings from the real dataset, presents robustness checks and discusses some caveats. Section 7 concludes. An online Appendix contains the basic economic model of optimal admissions (part A), some additional figures and 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 formuale for calculating the confidence intervals for threshold differences (part C).

2 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 will develop a corresponding econometric model incorporating unobserved heterogeneity, which can be taken to admissions data. The basic elements of the economic model are as follows.

Let W denote an applicant's pre-admission characteristics, observed by the university. Let $\phi(w)$ denote a w-type student's expected outcome (e.g., expected future GPA) if he/she enrols; and let $\alpha(w)$ denote the probability that a w-type student upon being offered admission eventually enrols. Let $c \in (0,1)$ be 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 those individuals whose $\phi(w) \geq \gamma$, where γ is chosen to satisfy the budget constraint. The key feature of the above rule is that γ

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

does not depend on covariates, and so the value of an applicant's W affects the decision on his/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 replaced 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 socio-demographic groups. The key challenge is to allow for the possibility that admission-tutors' 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. Therefore, we now turn to the task of constructing an econometric model incorporating unobserved heterogeneity in an empirical model of admissions.

2.1 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.

Now, let Z denote an index of academic ability of applicants, based on "soft" characteristics, such as evidence of enthusiasm, academic reference letters, etc., which are *unobservable*

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

to the analyst but observed by the admission-tutor. This may also include any random idiosyncrasies in the tutors' expectation formation process.[2] We assume that larger values of Z, without loss of generality, denote higher perceived academic potential.

Under meritocratic admissions, admission tutors 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 an 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,

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

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, females than males implies that a male applicant with the same expected outcome as a female applicant is more likely to be admitted. Conversely, under affirmative action type policies, γ will be lower for those demographics which 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 in our set-up, 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 females have a higher expected performance than males with identical X. This notion of fairness differs from one which requires that individuals who are identical on publicly verifiable variables (i.e., the Xs) 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.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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

3 Data

Our empirical analysis is based on admissions data for two recent cohorts of applicants to a competitive and popular undergraduate degree programme at a selective UK University. Students enter British universities to study a specific subject, rather than the US model of starting a general curriculum, followed by specialization in later years. Consequently, admissions are conducted primarily by faculty members (i.e., 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 where extra-curricular achievements are given no weight. In that sense, these admissions are more comparable with Ph.D. admissions in US universities. Furthermore, almost all UK applicants sit two common school-leaving examinations, viz., the 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, and 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 US, and write an essay that is graded.

Choice of sample: For our empirical analysis, we will focus on UK-domiciled applicants. The application process consists of an initial stage whereby a standardized "UCAS" form 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 and the rest rejected. Selected candidates are then assessed via a face-to-face interview and the interview scores

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

are recorded centrally. This sub-group 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 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.

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 0, below. 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.

Group identities G: We consider academic efficiency of admissions with regards to two different group identities, viz., type of school attended by the applicant and the applicant's gender. Selective universities in the UK 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 very much higher than in state schools (Graddy and Stevens, 2005), have an unfair advantage in the admission process. This is of special concern in a country like the UK where most selective universities are largely funded by the taxpayer. The issue of gender differences in admission and academic performance is, of course, a more universal issue. In the UK, as in most OECD countries, the higher education participation rate is higher for women, having overtaken that for men in 1993. However, selective universities in the UK have lagged behind the trend: in 2010-11, 55% of undergraduates across all UK universities were female, but 44% of students admitted to the

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

university we are analyzing were female. Typically, gender imbalances are more pronounced in certain programmes and includes the one we study, where male enrolment is nearly twice the female enrolment.

In our dataset, we can also match the post-admission academic performance of *admitted* students to their pre-admission 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 pre-admission 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 zero (c.f., Rothstein, 2004 for 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 which may only be positively correlated with eventual performance in 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 Fig. 3 and 6 below) as suggestive evidence of where our results might have arisen from.

4 Assumptions

In order to develop a test of meritocratic admissions, which 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 $G =$ '*school type*' and $X = (SAT, GPA, male)$, then $x_g \succeq_\varepsilon x_h$ means that applicant i and j are both male or both female and that $SAT_i > SAT_j + \varepsilon\sigma_{SAT}$

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

and $GPA_i > GPA_j + \varepsilon\sigma_{GPA}$, 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 τth quantile of the random variable Z given the random variable A.

Throughout the rest of the paper, we will 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,

$$\text{Median}\left[Z|X = x_g, G = g\right] \geq \text{Median}\left[Z|X = x_h, G = h\right],$$

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 (see immediately below):

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 stochastic dominates that of Z conditional on $X = x_h$, $G = h$:

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

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).

Discussion: 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 which 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) which admission

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

tutors view as positive determinants of future performance and hence, under the assumption of being academically motivated, would weigh positively in 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 below by providing evidence which strongly suggests 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 pre-admission 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 – viz., (i) when the distribution of the observable covariates are balanced across treatment and control groups in quasi-experimental designs, it is taken to imply that they are also balanced in terms of unobservables (e.g., Greenstone and Gayer, 2009) and (ii) orthogonality of an instrument with observed covariates is taken as suggestive evidence that it is orthogonal with unobserved covariates (e.g., Angrist and Evans, 1998, p. 458). In our context, the type 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 via reference 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 admis-

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

sion decisions, tutors are likely to observe these characteristics for current applicants via their dossiers 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 which are themselves continuously distributed.

Remark 2 *Note that assumption M/SD does* ***not*** *say that applicants with higher X have higher Z with probability one; 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, viz., $D = \mathbf{1}\{\phi(X, G, Z) \geq \gamma_G\}$. See Manski (1975) 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 as well as 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 under-represented communities.*

Empirical evidence of median-dominance: Among the pre-admission variables that we observe in our dataset, only the score on the interview is assigned by tutors. This is the type of variable most likely to be missing in other datasets 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

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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 =(GCSEscore, Aptitude_test_score, 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 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

$$\text{Median}[interview \mid X_g, g = male] - \text{Median}[interview \mid X_h, h = female]$$

across all paired realizations (X_g, X_h) satisfying $X_g \succeq_\varepsilon X_h$. We choose $\varepsilon = 0.0$; if we demonstrate median dominance for $\varepsilon = 0.0$, 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 $\varepsilon = 0$. In the appendix, we also show analogous histograms for the 25th and 75th quantiles with $\varepsilon = 0.0$. 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 evidence, we calculate the correlation matrix among the various indicators of academic merit at the pre-admission stage. These are reported in the online appendix, from where it is evident that all correlations are strictly

[4]Since we use only those applicants who were summoned for interview, there is an additional level of selection which can further weaken the correlation between unobservables and observables. Our "test" (c.f. Fig. 2, below) therefore assesses the extent of correlation remaining after both levels of selection.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

positive, which lends further support to assumption M/SD.

The evidence presented above 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 the above graph and tables 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 $x_g \succeq_\varepsilon x_h$, then $\phi(x_g, g, z) > \phi(x_h, h, z)$ for any z, and any $g \neq h$.

Discussion: 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 a 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 him 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 $x \succeq_\varepsilon x'$.

As a referee has pointed out, there is some evidence from the US state of California that females with lower SAT scores and high-school GPA than males have performed systemat-

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

ically better in college examinations (c.f. Leonard and Jiang, 1999, Rothstein, 2004). This does seem somewhat unlikely in our application, given Figure 1 above 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 $X_g^{con} \equiv \left(raw_score - \mu_g\right)/\sigma_g$, where μ_g and σ_g are the mean and standard deviation of the raw score *within group* g. Accordingly, the condition $X_{male}^{con} \succeq_\delta X_{female}^{con}$ refers to those male-female pairs where the males have higher *relative* scores than females, i.e., $\frac{X_{male}-\mu_{male}}{\sigma_{male}} \geq \frac{X_{female}-\mu_{female}}{\sigma_{female}} + \delta$. 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 $x_g^{con} \succeq_\delta x_h^{con}$, then $\phi\left(x_g^{con}, g, z\right) > \phi(x_h^{con}, h, z)$ for any z, and any $g \neq h$.

This assumption means that candidates whose performances are in the top echelons *of their own socio-demographic group*, will be perceived to be academically stronger. It thus allows for "biased" performance measures, e.g., that female applicants with lower raw scores on pre-entry evaluations may perform better in college exams, on average, and may therefore be favoured by admission officers over males 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 reported below, we report results for $\varepsilon = 0.1$ and $\varepsilon = 0.25$ to ensure that there is no overlap in observable

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

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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, below.

4.1 Identification Analysis

We show how assumption 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

$$\begin{aligned} p(x,g) &: = \Pr[D=1|X=x,G=g] \\ &: = \Pr\left[\phi(X,G,Z) > \gamma_g | X = x_g, G = g\right], \end{aligned}$$

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

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

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

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

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

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

Under the stronger SD assumption, non-emptiness of the set

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

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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

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

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 $x_g \succeq_\varepsilon x_h$ 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 farther 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 non-empty. If so, then one can conclude that $\gamma_g < \gamma_h$.

Remark 5 *The logical structure of our analysis is that if $S^{SD}(g, h, \varepsilon)$ is nonempty, then we can conclude that $\gamma_g > \gamma_h$. But it is possible that although $\gamma_g > \gamma_h$, we find that $S^{SD}(g, h, \varepsilon)$ 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*

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

assumptions justifying them, one cannot in general detect ***any*** *threshold difference – however large they might be.*

Alternative Identification Strategies: The above methodology may be contrasted with some alternative strategies proposed in the literature in non-educational contexts. For instance, in the context of healthcare, Chandra and Staiger (2009) attempt to identify difference 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 healthcare 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 (c.f. 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.

5 Estimation and Inference

Given the identification analysis above, 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 6 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), viz.,

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

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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

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

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 CLR (2013). Although our identification method is nonparametric in the sense of not requiring functional form specifications, estimation and inference for the nonparametric case is 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,

$$p(x_g,g) = \Pr\left[D=1|\,(X,G)=(x_g,g)\right] = \Phi\left(x_g'\boldsymbol{\delta}_{0,g}\right); \quad \text{and} \quad p(x_h,h) = \Phi\left(x_h'\boldsymbol{\delta}_{0,h}\right),$$

where $(\boldsymbol{\delta}_{0,g},\boldsymbol{\delta}_{0,h})$ are the probit coefficients; and Φ is the C.D.F. of the standard normal. Note that under our parametric specification, $\Phi(x_g'\boldsymbol{\delta}_g) \leq \Phi(x_h'\boldsymbol{\delta}_h)$ is equivalent to $x_g'\boldsymbol{\delta}_g \leq x_h'\boldsymbol{\delta}_h$ and thus

$$\mathcal{SD}(g,h,\varepsilon) = \left\{x_g \succeq_\varepsilon x_h,\ x_g'\boldsymbol{\delta}_{0,g} \leq x_h'\boldsymbol{\delta}_{0,h}\right\},$$

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

$$\theta_0 := \inf_{(x_g,x_h)\in\mathcal{X}_g\times\mathcal{X}_h,\ x_g\succeq_\varepsilon x_h} \left[x_g'\boldsymbol{\delta}_{0,g} - x_h'\boldsymbol{\delta}_{0,h}\right].$$

The quantity θ_0 is exactly of the form analyzed in CLR (2013). We construct a one-sided 95% confidence interval $\hat{C}_n(0.95) = \left(-\infty, \hat{\theta}_{n0}(0.95)\right)$ 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 non-empty.

6 Empirical Analysis

Summary statistics: We provide summary statistics for our sample in Table 1. The left half of table 1 shows that male applicants have better aptitude test scores and interview averages.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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 characteristics except for a slightly higher GCSE for the former.

In Table 2 we report the results of a probit regression of receiving an offer across all applicants. Table 2 strengthens the findings from Table 1 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% – which are about 10 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 2 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).

6.1 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 above.

In accordance with the discussion in Section 5, the first step is to examine emptiness of $\mathcal{SD}(g, h, \varepsilon)$ using data on only X and D. We first investigate this graphically in Figure 2 by

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

plotting the marginal C.D.F. of the difference in admission probabilities $p(X_g, g) - p(X_h, h)$ for pairs of (X_g, X_h) satisfying $X_g \succeq_\varepsilon X_h$ 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 emptiness of $\mathcal{SD}(g, h, \varepsilon)$. When the event $\{X_g \succeq_\varepsilon X_h\}$ happens with positive probability, an empty $\mathcal{SD}(g, h, \varepsilon)$ is equivalent to $\Pr[X_g \succeq_\varepsilon X_h,\ p(X_g, g) < p(X_h, h)] = 0$, where the probability is taken with respect to the distributions of X_g and X_h. Therefore, a positive mass at and below zero for these C.D.F.'s indicates that $\mathcal{SD}(g, h, \varepsilon)$ is nonempty. In the left panel, when $g = male$, $h = female$, the C.D.F. is represented by the solid curve labelled male_fem; and when $g = female$ and $h = male$, it is the dashed curve, labelled fem_male. A positive height at zero indicates that applicants with higher observables in the first group has lower admission probabilities than the second.

Clearly, the first curve has significant mass below zero and the dashed curve has almost no mass below zero, suggesting a positive probability that $p(X_{male}, male) < p(X_{female}, female)$ although $X_{male} \succeq_\varepsilon X_{female}$. 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 3, 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 5. We report results for $\varepsilon = 0.1$ (recall that we concluded dominance even with $\varepsilon = 0$, c.f., fig. 2). A negative upper limit indicates that the set $\mathcal{SD}(g, h, \varepsilon)$ is nonempty and consequently we reject the null of $\gamma_g \leq \gamma_h$ in favour of $\gamma_g > \gamma_h$. It is evident from the first four rows of table 3 that we reject emptiness for $g = male$, $h = female$ and for $g = indep$, $h = state$ but do not reject emptiness in the other cases. This suggests that males and private school applicants face higher admission thresholds. The exact upper limits of confidence intervals reported above vary slightly across functional specifications (e.g., 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

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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 schooltype. Results reported in the last two rows of table 3 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.

Interpretation of the empirical findings: It would be natural to conjecture that the threshold differences arise primarily from the implicit or explicit practice of affirmative action, viz., the overweighting of outcomes for historically disadvantaged groups. A second possibility is that, in face of political and/or media pressure, admission tutors try to equate an application success rate for, say, males with one for females, which is also consistent with our empirical findings (see Tables 1 and 2). 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 are set a lower admission threshold in order to encourage more female candidates to apply in future. Note from Table 1 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 C.D.F.s 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/schooltype for en-

[6] As 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, c.f. Goldin, et al, 2006 and Niederle and 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.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

rolled students. The regression output appears in table 4. The estimated CDFs of predicted performance by gender and schooltype are plotted in Figure 3B.

It is clear that in both graphs, the male distribution first-order stochastic 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 3 imply that allowing for unobserved heterogeneity does not change this scenario substantively, and suggests that equating the application success-rates (see table 1) leads to the use of higher admission thresholds for male and, to a lesser extent, for 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, i.e., a "high second class" mark of 64% – a minimum requirement for entry into most postgraduate programmes – as the relevant outcome produces a very similar result, presented in Fig. 6.

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 cut-off across G, thereby generating insignificant effects of gender and school-type on admission probabilities, both unconditionally (c.f. Table 1) and conditionally on past test-scores (c.f. Table 2). Such behavior could arise either from an institutional norm banning any conditioning on demographic characteristics, or from the tutors' belief that such characteristics have no explanatory power beyond the pre-admission 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 stu-

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

dents are more able than independent school students with the same test score, and therefore perform better in post-admission exams. If tutors ignore G, then an independent and a state school student with identical pre-admission test scores will have 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 favour of type-g applicants and award them higher interview marks (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 candidates was systematically awarded higher interview scores relative to their performance in the other more "objective" tests. For example for $g = male$ and $h = female$, if males 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.

6.2 Some Robustness Checks

Biased test-scores: One feature of our approach is that we are taking the pre-admission test scores as true indicators of academic merit. However, students from privileged backgrounds might perform well in these tests simply on account of having being coached extensively. It is not possible to conduct any analysis of meritocracy if no previous measure of achievement can be taken to be accurate. As mentioned above, post-admission performance is not observed for non-admitted 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. Accordingly, we repeat the above analysis by replacing *each* test-score by its standardized version and invoking assumption CM', above. Recall the condition $X_{male\delta} \succeq_{\delta} X_{female}$ which refers to those male-female pairs where the males have higher *relative* scores than females. Then we can conclude that group g faces a higher

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

threshold than group h if $\theta_0 < 0$, where

$$\theta_0 \equiv \inf_{(x_g,x_h)\in\mathcal{X}_g\times\mathcal{X}_h,\ x_g\succeq_\delta x_h} \left[x_g'\boldsymbol{\delta}_{0,g} - x_h'\boldsymbol{\delta}_{0,h}\right].$$

The results from this exercise are shown in Table 3, in the last column titled "Standardized Scores" corresponding to $\delta = 1.25$ (the smallest δ for which histograms analogous to those in Figure 1, above, 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 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 3, 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 schooltype 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-third are interviewed. Figure 5 presents suggestive evidence regarding 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 Fig. 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

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

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.

6.3 Caveats

Several caveats apply to our methods and data. The first is that we ignore peer-effects, both at the individual level and also at the institutional level. For example, it is possible that an applicant is admitted (or rejected) because he/she is deemed to have the potential to create positive (negative) externalities on his/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 get 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 6 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 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 US), it has been found that female students perform better in college exams than males with the same pre-admission test-scores. If that were true, admission officers may admit female applicants who have scored relatively lower on pre-admission assessments. This is unlikely to be the case in

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

our application; indeed, Fig 3A shows that post-entry college performance of males first-order stochastic dominates that of females, which is inconsistent with the superior female performance explanation. Moreover, Fig 3B shows that predicted college performance on the basis of observables is also stochastically higher for males, which provides further evidence against that explanation. However, when applying our methods to other contexts, it would be advisable to draw graphs analogous to Fig 3A and 3B as a preliminary check.

7 Summary and Conclusion

This paper has proposed an empirical approach to testing, on the basis of micro-data, 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 UK university, we find that admission thresholds faced by male applicants are significantly higher than females while those for private-school applicants 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 6.3. Beyond the application to college-admissions, our methods are potentially useful for testing fairness of other binary decisions such as loan-approval, surgery-referrals etc., where allegations of bias are common.

References

[1] Arcidiacono, Peter (2005) Affirmative action in higher education: How do admission and financial aid rules affect future earnings?, Econometrica, 73-5, 1477-1524.

[2] Arcidiacono, Peter., and Mike Lovenheim (2015). Affirmative action and the quality-fit trade-off. No. w20962. National Bureau of Economic Research.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

[3] Becker, Gary. (1957) The economics of discrimination, University of Chicago Press.

[4] Bertrand, Marianne, Remma Hanna & Sendhil Mullainathan (2010) Affirmative action in education: Evidence from engineering college admissions in India, Journal of Public Economics, 94, 1-2, 16-29.

[5] Bhattacharya, Debopam. & Pascaline Dupas (2012) Inferring efficient treatment assignment under budget constraints, Journal of Econometrics, 167, 168-196.

[6] Bhattacharya, Debopam (2013) Evaluating treatment protocols using data combination, Journal of Econometrics, 173, 160-174.

[7] Card, David & Alan Krueger (2005) Would the elimination of affirmative action affect highly qualified minority applicants? Evidence from California and Texas, Industrial and Labor Relations Review, 58-3, 416-434.

[8] Chandra, Amitabh & Doug Staiger (2009) Identifying provider prejudice in medical care, Mimeo, Harvard University and Dartmouth College.

[9] Chernozhukov, Victor, Simon Lee & Adam Rosen (2013) Intersection bounds: Estimation and inference, Econometrica, 81-2, 667-737.

[10] Espenshade, Thomas et al (2004). Admission Preferences for Minority Students, Athletes, and Legacies at Elite Universities. Social Science Quarterly 85.5 (2004): 1422-1446.

[11] Fryer Ronald. & Glenn Loury (2005) Affirmative action and Its mythology, Journal of Economic Perspectives, 19-3, 147-162.

[12] Goldin, Claudia, Larry Katz, and Ilyana Kuziemko. 2006. The Homecoming of American College Women: The Reversal of the College Gender Gap. Journal of Economic Perspectives, 20(4): 133–56.

[13] Graddy, Kathleen & Margaret Stevens (2005) The Impact of School Inputs on Student Performance: An Empirical Study of Private Schools in the United Kingdom, Industrial and Labor Relations Review, 58-3, 435-451.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

[14] Greenstone, Michael & Tom Gayer (2001) Quasi-experimental and experimental approaches to environmental economics, Journal of Environmental Economics and Management, 57, 21-44.

[15] Heckman, James (1998) Detecting discrimination, Journal of Economic Perspectives, 12-2, 101-116.

[16] Hinrichs, Peter. "The effects of affirmative action bans on college enrollment, educational attainment, and the demographic composition of universities." Review of Economics and Statistics 94.3 (2012): 712-722.

[17] Hoxby, Caroline (2009) The changing selectivity of American colleges, Journal of Economic Perspectives, American Economic Association, 23-4, 95-118.

[18] Hurwitz, Michael (2011). The impact of legacy status on undergraduate admissions at elite colleges and universities, Economics of Education Review, vol. 30, issue 3, pages 480-492.

[19] Kane, Thomas (1998) Racial and ethnic preference in college admissions, in Christopher Jencks and Meredith Phillips (eds.), The Black-White Test Score Gap, Washington: Brookings Institution.

[20] Keith, Simon (1985) Effects of affirmative action in medical schools – A study of the class of 1975, The New England Journal of Medicine, 313, 1519-1525.

[21] Kobrin, Jonathan (2008) Validity of the SAT for predicting first-year college grade point average, College Board, New York.

[22] Kuncel, N (2001) A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. Psychological Bulletin, 127, 162-181.

[23] Leonard, David, and J. Jiang (1999). "Gender bias and the college predictions of the SATs: A cry of despair." Research in Higher Education 40.4: 375-407.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

[24] Manski, Charles (1988) Identification of binary response models, Journal of the American Statistical Association, 83, 729-738.

[25] Manski, Charles (2009): Identification for Prediction and Decision, Cambridge, Massachusetts: Harvard University Press.

[26] Niederle, Muriel, and Lise Vesterlund (2010). Explaining the gender gap in math test scores: The role of competition. The Journal of Economic Perspectives, 129-144.

[27] Ogg , Thomas (2009) Schooling effects on degree performance: a comparison of the predictive validity of aptitude testing and secondary school grades at Oxford University, British Educational Research Journal, 35-5.

[28] Persico, Nicola (2009) Racial profiling? Detecting bias using statistical evidence, Annual Review of Economics, 1, 229-254.

[29] Rothstein, Jesse (2004). College performance predictions and the SAT. Journal of Econometrics, Elsevier, vol. 121(1-2), pages 297-317.

[30] Sackett, Paul (2009) Socioeconomic status and the relationship between the SAT and freshman GPA - An analysis of data from 41 colleges and universities, available online at:
http://professionals.collegeboard.com/data-reports-research/cb/SES-SAT-FreshmanGPA

[31] Sander, Richard (2004): A Systemic Analysis of Affirmative Action in American Law Schools, 57 Stanford Law Review 367-483

[32] Sawyer, Richard (2010) Usefulness of high school average and ACT scores in making college admission decisions, available online at:
http://www.act.org/research/researchers/reports/pdf/ACT_RR2010-2.pdf

[33] Zimdars, Anna & Anthony Heath (2009) Elite higher education admissions in the arts and sciences: Is cultural capital the key?, Sociology, 4, 648-66.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

Table 0: 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

Note: 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/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.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

Table 1: Means by Gender and by Schooltype

Variable	**Female (N=241)**	**Male (N=394)**	**pvalue_diff**	**State (N=355)**	**Indep (N=280)**	**pvalue_diff**
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

Note: The data pertain to two cohorts of applicants. The variable names are explained in table 0. 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 zero at 5%.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

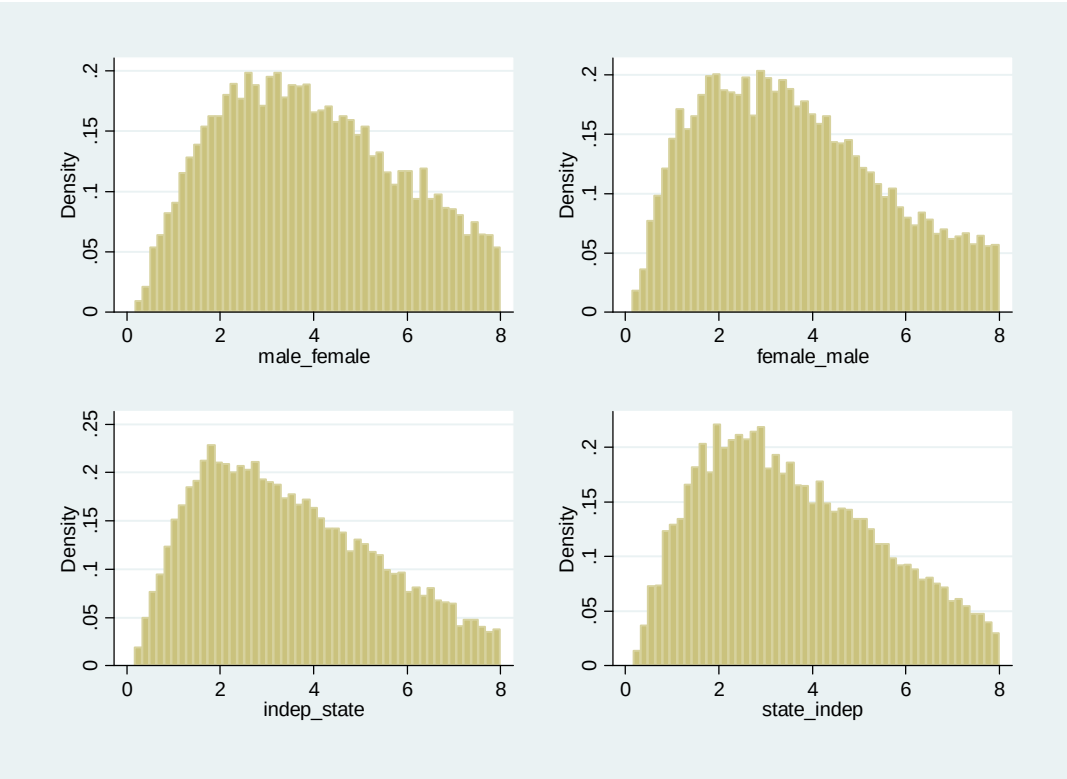
Table 2: Probit Regression of Admission

Variable	Coef.	Coef/std.err	Marginal Effect	Marg.Eff/Std.Err
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

Note: Probit regression of eventual admission for all UK-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 UK-based applicants. Marginal effects are calculated at mean values of covariates and for moving from 0 to 1 for male and independent. Gender and schooltype remain insignificant (highlighted in yellow) even after controlling for past test-scores.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

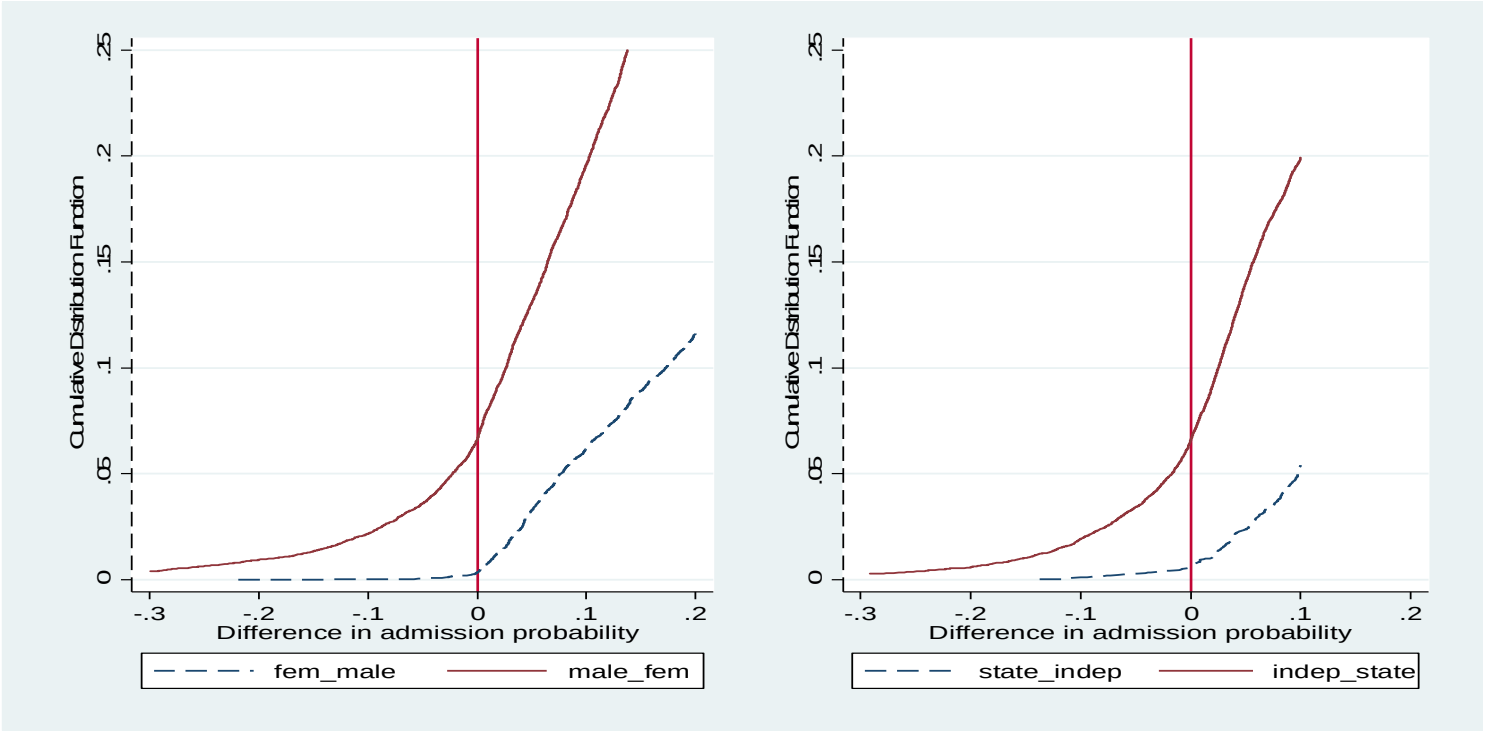
Figure 1: Evidence of Median Dominance



Note: 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: Median[interview | Xmale,G=male]-Median[interview | Xfemale,G=female] across all paired realizations (Xmale,Xfemale) satisfying Xmale⩾Xfemale.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

Figure 2: Graphical evidence of different admission thresholds



Note: The above graphs plot the marginal C.D.F. of the difference in admission probabilities p(Xg,g)-p(Xh,h) for pairs of (Xg,Xh) satisfying Xg>ε Xh for ε=0.1 for various combinations of g and h. A positive height at zero 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 males with higher observables have lower admission probability than a subgroup of females with lower observables.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

Table 3: Testing Unequal Thresholds

Difference	**ε=0.1**	**ε=0.25**	Quadratics in Pre-Admission Scores **ε=0.1**	Standardized scores **δ=1.5**
g=male, h=female	-1.73	-2.02	-3.49	-2.01
g=female, h=male	0.57	0.67	0.684	0.43
g=indep, h=state	-1.29	-0.58	-2.75	0.012
g=state, h=indep	0.92	0.04	0.635	1.87
g=state_male, h=state_female	-1.36	-1.01	-6.85	-1.19
g=indep_male,h=indep_female	-1.11	-3.39	-2.7	-3.56

Note: 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 ε=0.1 and ε=0.25 correspond to evaluating difference in admission probability (as a function of gcsescore, 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 pre-entry 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 sections 8.2 and 8.5, respectively. The last-but-one column shows the results when qudratics and second-order interactions between all pre-admission performance measures are used as additional controls, beyond the linear versions of them, to predict admission probabilities, as a robustness check.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

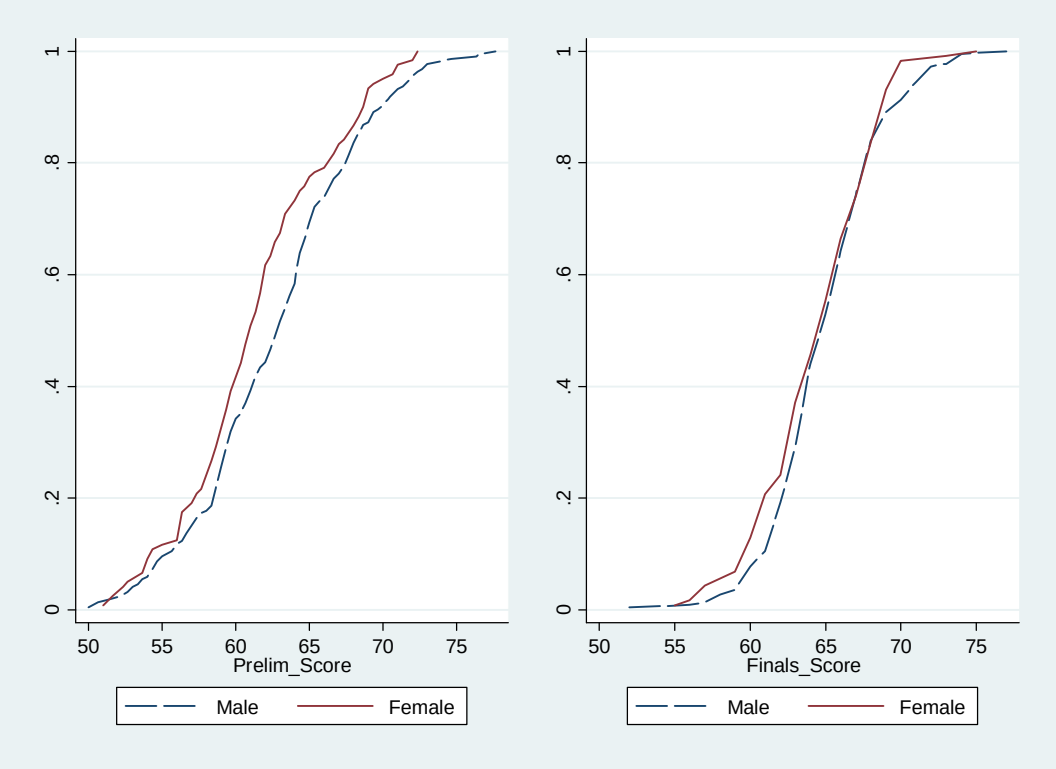
Table 4: Regression of first year performance on observable covariates

	Coefficient	Std error	t-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	0.69	1.66
indep	0.41	0.68	0.75

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

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

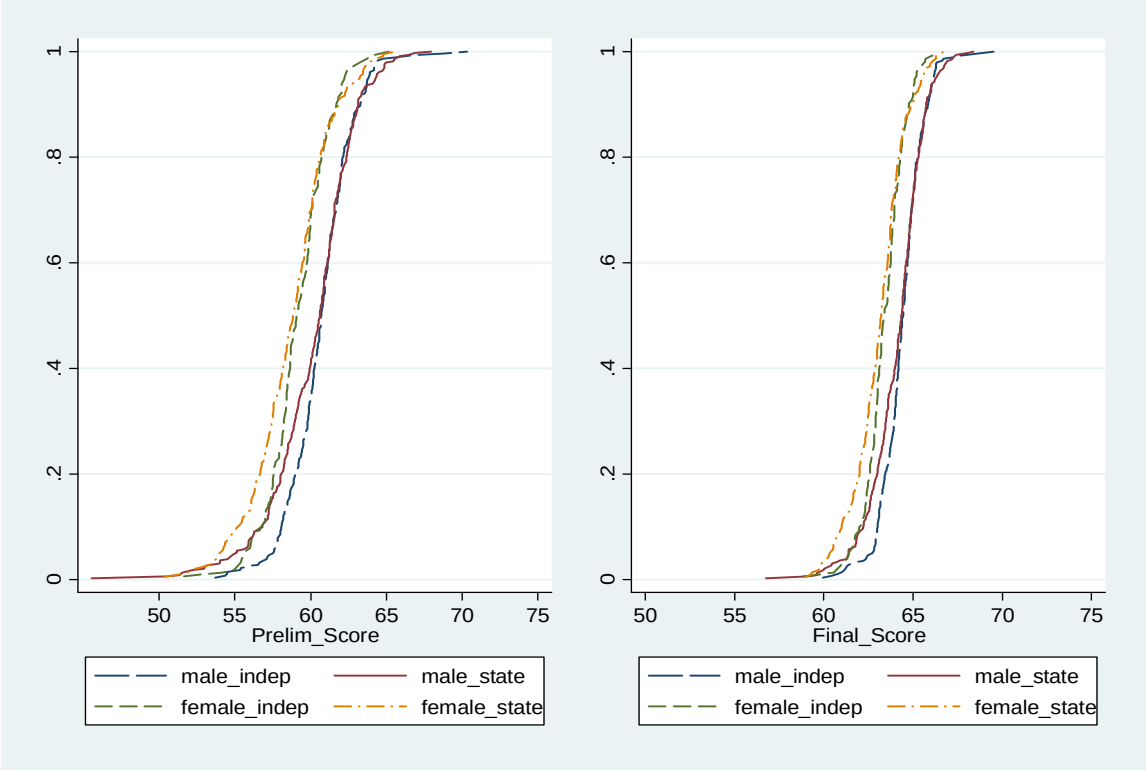
Figure 3A: CDF of first year (prelim) and third year (final) performance, by gender



Note: CDF of first-year (left panel) and final year (right panel) performance in college for admitted candidates. The male CDFs are seen to lie almost entirely to the right of the female CDFs, with dominance more pronounced for prelims.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

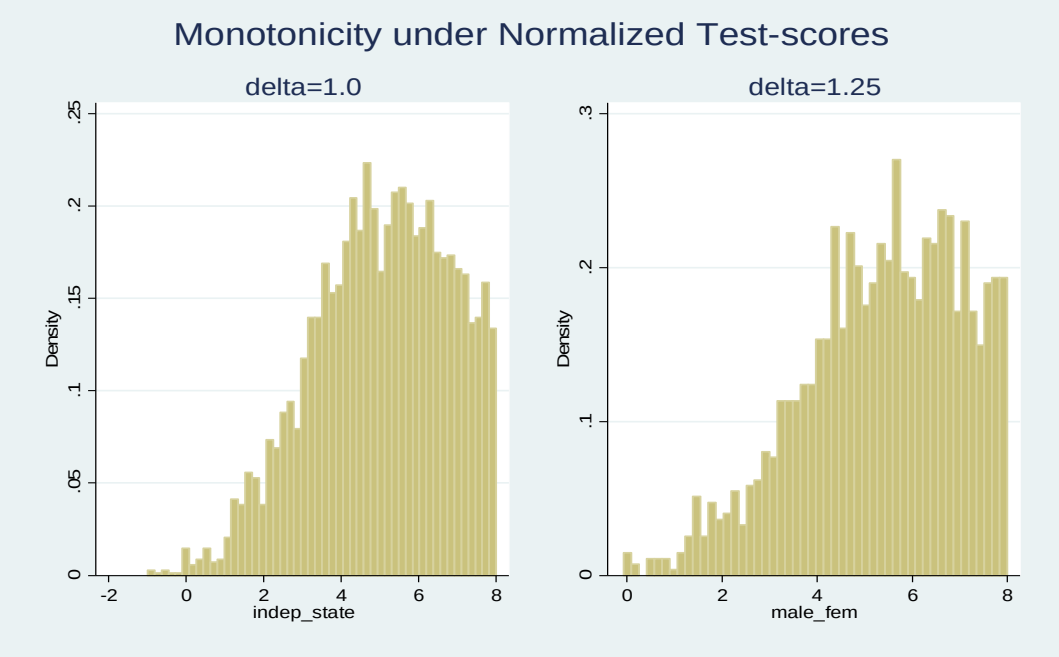
Figure 3B: CDF of predicted first year (prelim) and third year (final) performance based on observables, by gender and school-type



Note: CDF of predicted first-year (left panel) and final year (right panel) performance in college 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.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

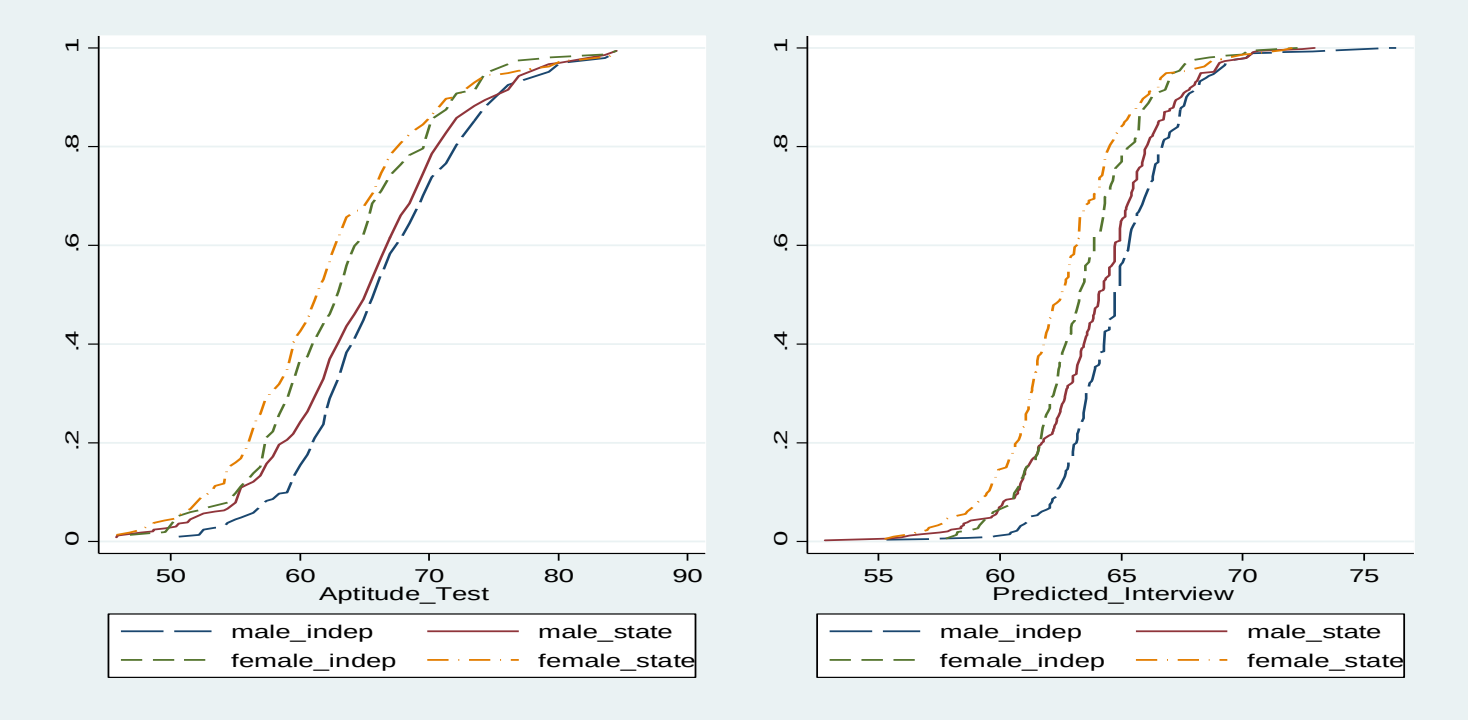
Figure 4: Testing Monotonicity of Median Interview Score in "Contextualised" Test-scores



Note: The above graphs plot histograms of the difference in admission probabilities p(Xmale,male) - p(Xfemale, female) for pairs of (Xmale, Xfemale) satisfying Xmale>Xfemale+ δ, for δ =1.0 and δ =1.25, where Xmale, Xfemale are the *standardized* test-scores observed prior to admissions. The smallest δ for which these histograms have positive support is δ =1.25. We use this value of δ to do our robustness checks, as explained in the paper in Section 8.3.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

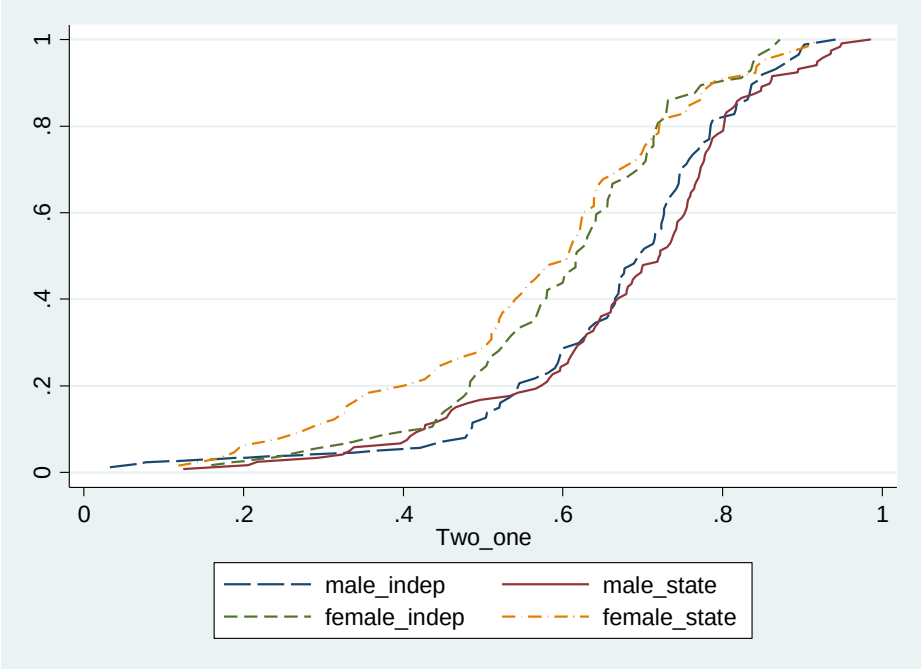
Figure 5: First-stage Selection



Note: The above graphs present suggestive evidence regarding first-stage selection of candidates. The left panel plots the CDF of the raw aptitude test-scores for those making it to the interview stage. The right graph plots the CDF of predicted interview scores based on aptitude test score and GCSE score, and is analogous to Figure 3 above which pertains to the second stage of selection. A common success rate across gender and schooltype 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.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

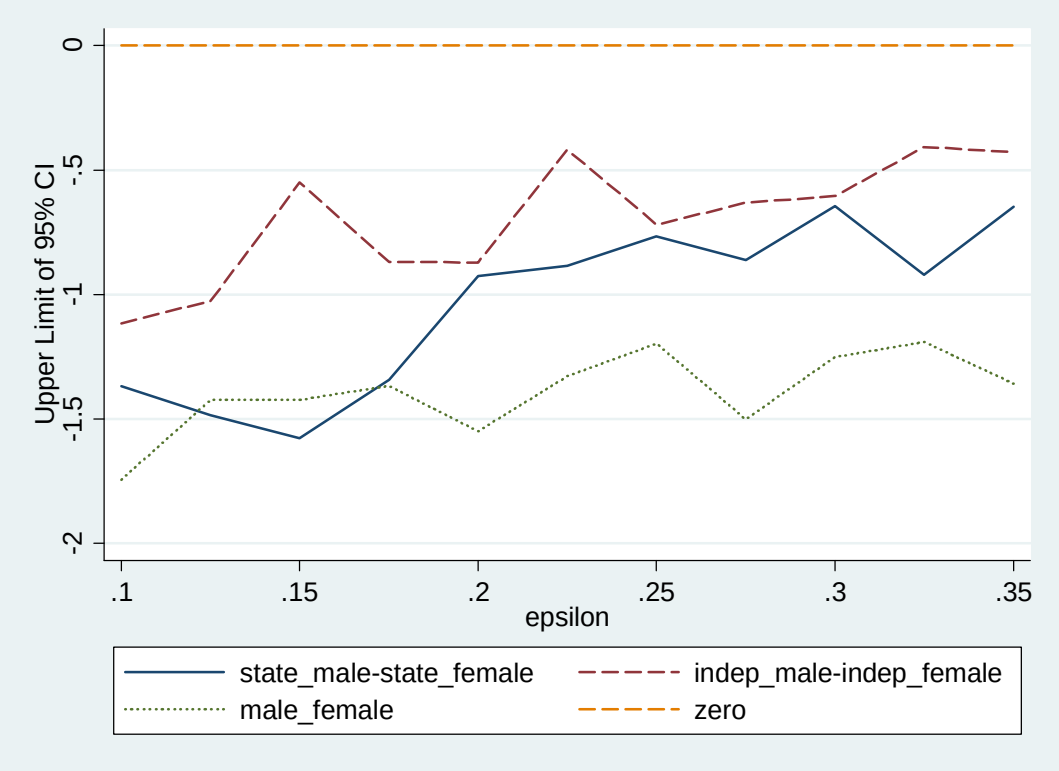
Figure 6: Predicted Probability of attaining a 2.1 level mark



Note: The above graph plots the CDF of the predicted probability of getting at least a high second class level mark (64%) in the 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 schooltype 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, above.

Review of Economics and Statistics Just Accepted MS.
doi:10.1162/REST_a_00618
© by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

Figure 7: Effect of ε on gender-gap in admission thresholds



Note: In this figure, we examine how the overall male-female gap in thresholds differs by school-type, and also 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.