Are University Admissions Academically Fair?

Debopam Bhattacharya[†,*], Shin Kanaya[‡] and Margaret Stevens[†]

[†]University of Oxford and [‡]University of Aarhus

December 22, 2013.

Abstract: We develop a test of whether selective universities admit applicants with the highest academic potential and a measure of "nonacademic-bias" when they don't. Efficient admissions should equate the expected future performance of marginal candidates – the admission-threshold – across demographic groups but such thresholds are difficult to calculate due to unobserved characteristics. We assume that applicants who are better-qualified on standard observable indicators would on average, but not necessarily with certainty, appear academically stronger to admission-tutors based on characteristics observable to them but not us. This assumption yields informative bounds on *differences* in admission standards faced by different demographic groups, which are robust to omitted-characteristics problems. An application to admissions-data at a selective British university, using blindly-marked, future exam-performance as potential outcome, shows that males face significantly higher admission-standards and private school applicants less so. In contrast, application success-rates are equal across both gender and school-type.

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

1 Introduction

Admission practices at selective universities generate considerable public interest and political controversy, owing to their implications for socioeconomic mobility. For example, in the UK a highly publicized 2011 Sutton Trust report shows that nationally just 3% of schools – mostly expensive, independent (i.e., private) institutions – account for 32% of undergraduate admissions to Oxford

[*]Address for correspondence: Debopam Bhattacharya, Department of Economics, University of Oxford. Manor Road Building, Manor Road, OX1 3UQ, United Kingdom. Email: debobhatta@gmail.com

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's public universities and race-based affirmative action in American state-funded colleges have been the subject of intense controversy, the latter recently re-surfacing in the high-profile "Fisher versus University of Texas" lawsuit. In this context, it is of significant policy interest to measure the extent of equity-efficiency trade-off implicit in current admission protocols, based on micro-level admissions data. In this paper, we develop a rigorous empirical methodology to model such trade-offs and use it to devise a test for whether all applicants are held to the same academic standard during admissions.

Our approach is based on the productivity based view of optimal decisions, in the tradition of Becker (1957), and focuses on the expected future performance of university-entrants. Viewed in this light, if admissions are purely meritocratic, then the marginal admitted student from a state-school should be expected to perform just as well as the marginal admit from a private school. But her expected performance would be worse if affirmative action leads to admitting state-school students who are not expected to perform at or above the same standard as marginal private school students in future exams. The difference between expected performances of marginal candidates across demographic groups can therefore be interpreted as a direct measure of efficiency loss. A challenge in implementing this approach is that a researcher typically observes a subset of the relevant applicant characteristics used by admissions-tutors and the distributions of the unobserved characteristics may – and usually do – differ across groups. Such "omitted characteristics" problems jeopardize the researcher's attempt at reconstructing the decision-maker's perceptions and make it hard to assess whether the decision-maker acted in an academically unbiased way. This type of problem has been recognized by previous researchers, especially in the context of labor market hiring; see, for instance, Heckman (1998), Blank et al. (2004) and the references therein. In the present paper, we use methods from the recent econometric literature on partial identification analysis to devise a test for academically fair admissions, based on the *differences* in admission-thresholds faced by different demographic groups which are 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 provide empirical evidence – is that

applicants who are significantly better in terms of easily observable indicators of academic potential should statistically – not 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 under this assumption, one can identify a lower bound on the magnitude of the threshold *differences* in admission thresholds applied to different demographic groups. We apply these methods to analyze admissions data from a popular undergraduate programme of study at a selective UK University, focusing on future academic performance as the potential outcome of interest. In our sample, the application success rates are almost identical across gender and type of school attended by the candidate (an "independent" school being an indicator of higher socioeconomic status), both before and after controlling for key covariates. However, applying our method of threshold detection, we find that admission standards faced by applicants who are male or from independent schools exceed those faced by females or state school applicants. This finding is suggestive of some degree of affirmative action – either explicit or implicit – within the admission process, which is not apparent from the equal success rates, thereby illustrating the usefulness of our approach. We also find evidence suggestive of "catch up", whereby the performance gap in the third year final examinations between marginal candidates who are female or from state-schools and marginal candidates who are male or from independent schools appears to be smaller than that in first year performance.

Related literature: A large volume of research exists in educational statistics on the analysis of admissions to selective colleges and universities, focusing mainly on the United States. For a broad, historical perspective on selectivity in US college admission, see Hoxby (2009). We are not aware of any previous attempt in the academic literature in education, economics or applied statistics to formally test *outcome-oriented* efficiency – in Becker's sense – of college admissions. A distinguishing feature of the present paper is that it focuses on the predicted eventual outcomes of the students and thereby shows that equal success rate in admissions across demographic groups can be consistent with very different admission standards across these different groups. Indeed, that is precisely what we conclude in our empirical application. These conclusions are based on the outcome of the *marginal* admits in different demographic groups, which is in contrast to many other studies – both academic and policy-oriented – which compare either *average* pre-admission test-scores (c.f. Zimdars et al., 2009, Herrnstein and Murray, 1994) or *average* post-admission performance across *all* admitted students from different socioeconomic groups (c.f. Keith et al., 1985, Sackett et al., 2009, Kane and William, 1998). To our knowledge, the only other work in

this literature which focuses on marginal admits is Bertrand, Hanna and Mullainathan (2010), who examined the consequences of affirmative action in admission to an Indian college. In their setting, admission was based on score in a single entrance exam; admission thresholds differed by applicants' social caste and were publicly announced. This 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 methodology is designed to deal with this more general scenario.

In Economics, our paper 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 and a comprehensive bibliography. On the empirical end, Arcidiacono (2005) uses a structural model of admissions to simulate the potential, counterfactual consequences of removing affirmative action in US college admission and financial aid on applicant earning, while Card and Krueger (2005) describe the reduced-form impact of eliminating affirmative action on minority students' application behavior in California. In contrast to these works, the present paper may be viewed as one that attempts to *detect* the presence and quantify the extent of affirmative action in prevalent admission practises, based on admissions-related micro-data.

The rest of the paper is organized as follows: Section 2 sets up a simple theoretical model; Section 3 the corresponding empirical model of meritocratic admissions; Section 4 contains the identification analysis; Section 5 discusses inference; Section 6 discusses the data setting and reports a simulation exercise based on it; Section 7 reports the empirical findings and some robustness checks regarding the interpretation of the results; and Section 8 concludes. Technical proofs are collected in an Appendix.

2 Benchmark Optimization Model

We start by laying out a benchmark economic model of admissions to help fix ideas. Based on this economic model, in the next section we develop a corresponding econometric model incorporating unobserved heterogeneity, which can be taken to admissions data.

Let W denote an applicant's pre-admission characteristics, observed by the university. We 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. Also, let Y denote the applicant's future academic performance if admitted to the university (assumed to take on non-negative values, e.g., GPA), and the binary indicator D denote whether the applicant received an admission offer and the binary indicator A denote whether the admission offer was accepted by the applicant.

Let $\mathcal{W}$ denote the support of W, $F_W(\cdot)$ denote the marginal cumulative distribution function (C.D.F.) of W; $\mu^*(w)$ denote a w-type student's expected performance ($w \in \mathcal{W}$) 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 a constant denoting the fraction of applicants who are to be admitted, given the number of available spaces.

Admission protocols: We define an admission protocol as a probability $p(\cdot) : \mathcal{W} \to [0,1]$ such that an applicant with characteristics w is offered admission with probability $p(w)$. A generic objective of the university may be described as

$$\sup_{p(\cdot)\in\mathcal{F}} \int_{w\in\mathcal{W}} p(w)\, h(w)\, \alpha(w)\, \mu^*(w)\, dF_W(w) \quad \text{subject to} \quad \int_{w\in\mathcal{W}} p(w)\, \alpha(w)\, dF_W(w) \leq c.$$

Here, $\mathcal{F}$ denotes the set of all possible p's, and $h(w)$ denotes a non-negative welfare weight, capturing how much the outcome of a w-type applicant is worth to the university. For affirmative action policies, $h(\cdot)$ will be larger for applicants from disadvantaged socioeconomic backgrounds or under-represented demographic groups. The overall objective is thus to maximize total welfare-weighted outcome among the admitted applicants, subject to a capacity constraint. The solution to the above problem takes the form described below in Proposition 1, which holds under the following condition:

Condition C: $h(w) > 0$ and $\alpha(w) > 0$ for any $w \in \mathcal{W}$.[1] Further, for some $\delta > 0$,

$$\int_{w\in\mathcal{W}} \alpha(w)\, \mathbf{1}\{\mu^*(w) \geq 0\}\, dF_W(w) \geq c + \delta,$$

i.e., admitting everyone with $\mu^*(w) \geq 0$ will exceed the capacity in expectation.

Proposition 1 *Under Condition C, the solution to the problem:*

$$\sup_{p(\cdot)\in\mathcal{F}} \int_{w\in\mathcal{W}} p(w)\, h(w)\, \alpha(w)\, \mu^*(w)\, dF_W(w) \quad \textit{subject to} \quad \int_{w\in\mathcal{W}} p(w)\, \alpha(w)\, dF_W(w) \leq c$$

[1] Alternatively, we can simply redefine $\mathcal{W}$ to be the subset of the support of W with $\alpha(w) > 0$.

takes the form:

$$p^{opt}(w) = \begin{cases} 1 \text{ if } \beta(w) > \gamma; \\ q \text{ if } \beta(w) = \gamma; \\ 0 \text{ if } \beta(w) < \gamma, \end{cases} \tag{1}$$

where

$$\beta(w) := h(w)\mu^*(w); \quad \gamma := \inf\{r : \int_{w\in\mathcal{W}} \alpha(w)\mathbf{1}\{\beta(w) > r\} dF_W(w) \leq c\};$$

and $q \in [0,1]$ *satisfies*

$$\int_{w\in\mathcal{W}} \alpha(w)[\mathbf{1}\{\beta(w) > \gamma\} + q\mathbf{1}\{\beta(w) = \gamma\}] dF_W(w) = c.$$

The solution (1) is unique in the F_W*-almost-everywhere sense (i.e., if there is another solution, it differs from (1) only on sets whose probabilities are zero with respect to* F_W*).*

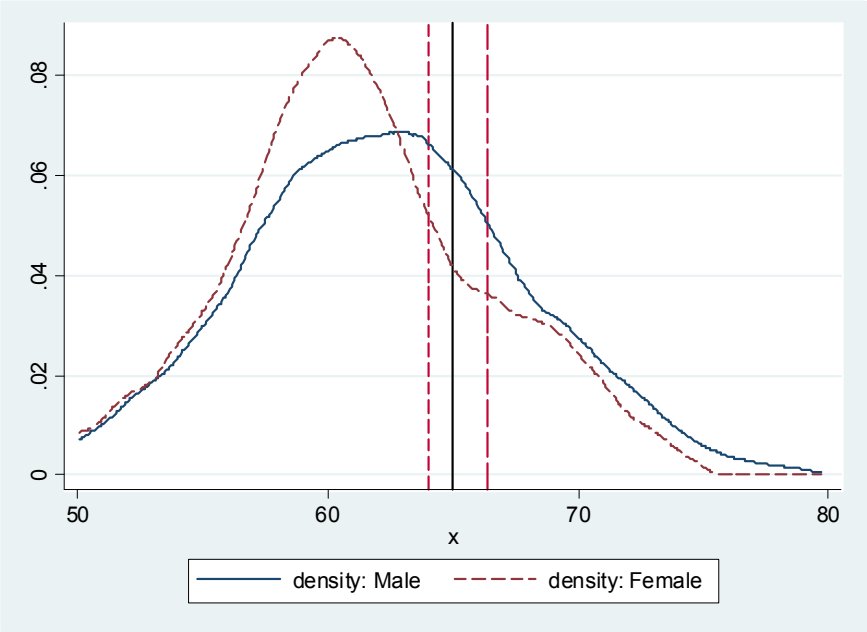
The result basically says that the planner should order individuals by their values of $\beta(W)$ and first admit applicants with those values of W for which $\beta(W)$ is the largest, then to those for whom it is the next largest and so on till all places are filled. If the distribution of $\beta(W)$ has point masses, then there could be a tie at the margin, which is then broken by randomization (hence the probability q). In the absence of any point masses in the distribution of $\beta(W)$, the optimal protocol is of a simple threshold-crossing form $p^{opt}(w) = \mathbf{1}\{\beta(w) \geq \gamma\}$. For the rest of the paper, we will assume that this is the case. It is useful to note that $\alpha(w)$ affects the admission rule only through its impact on γ; the intuition is that individuals who do not accept an offer of admission contribute nothing to the budget constraint and this is taken into account in the admission process.

Academically efficient admissions: We define an academically efficient admission protocol as one which maximizes total performance of the incoming cohort subject to the restriction on the number of vacant places. Such an objective is also "academically fair" in the sense that the expected performance criterion gives equal weight to the *outcomes* of all applicants, regardless of their value of W, i.e., $h(w)$ is a constant. In this case, the previous solution takes the form $p^{opt}(w) = \mathbf{1}\{\mu^*(w) \geq \gamma\}$, where γ solves

$$c = \int_{w\in\mathcal{W}} \alpha(w)\mathbf{1}\{\mu^*(w) \geq \gamma\} dF_W(w).$$

The key feature of the above rule is that γ does not depend on W and so the value of an applicant's W affects the decision on his/her application only through its effect on $\mu^*(W)$. 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. The following graph illustrates the idea.



Equal threshold versus equal admission rate

In this graph, the solid curve represents the marginal density of predicted future performance for males and the dotted curve that for females. Under identical thresholds, marked by the solid vertical line, the probability of acceptance equals the area – to the right of the line – under the solid density curve for male applicants and under the dotted density curve for female applicants. The graph shows that the latter area is significantly smaller, suggesting that if a common threshold were used, admission rate for female applicants would be lower. Conversely, equating admission probabilities across gender requires employing a larger threshold (marked by long dash) for males than for females (smaller dash). The difference between the thresholds is then a logical measure of deviation from meritocratic admissions. Indeed, if the density curves have identical right tails, then equal thresholds can be consistent with equal admission rates. 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 the predictions were based on more

characteristics than we the researchers observe, so that we cannot replicate the two density curves as in the previous graph.

3 Econometric Model

To set up our empirical framework, we assume that we observe the covariates X, G and the binary admission outcome D (= 1 if admitted, and = 0 otherwise) for applicants in the current year. In addition, we have data on several cohorts of applicants in past years who had enrolled in the university. For each such enrolled applicant, we observe X, G and the outcome of interest Y (e.g., examination score in the university). When referring to variables from past years or expectations calculated on the basis of past variables, we will use the superscript "^{P}". Thus, our aim is to evaluate academic efficiency of current year's admission, given data on (X, G, D) for all current year applicants and $(Y^P, X^P, G^P \mid A^P = 1)$ for past years' (successful) applicants, where $A^P = 1$ denotes having enrolled in the university. For later use, define

$$\mu^P(x, g) = E\left[Y^P | X^P = x, G^P = g, A^P = 1\right], \tag{2}$$

the conditional expectation of outcome Y^P for a past enrolled applicant given his/her characteristics $(X^P, G^P) = (x, g)$. Let $\mathcal{X}_g$, $\mathcal{X}_h$ denote the support of X for applicants of type g and h, respectively in the current year. Also, let $\mathcal{X}_g^P$ denote the values of X^P which occur among the admits of type g in past years and so one can estimate the values of $\mu^P(x, g)$ when $x \in \mathcal{X}_g^P$.

Now, let Z denote a scalar index of academic ability of a current applicant, based on characteristics which are *unobservable* to the analyst but observed by the admission-tutor. This may also include any random idiosyncrasies in the tutors' expectation formation process. 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 $\mu_i^* \equiv \phi(X_i, G_i, Z_i)$, their subjective assessment of how applicant i will perform when admitted. In accordance with our economic model, we assume that a current year applicant $i(\in \{1, \ldots, n\})$ 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 $\mu_i^* = \phi(x, g, z) \geq \gamma_g$, where μ_i^* denotes the subjective conditional expectation of applicant i's future performance calculated by the admission-tutor handling his file and γ_g denotes the university-wide baseline threshold for applicants of demographic type g. That is,

$$D_i = \begin{cases} 1 & \text{if } \phi(X_i, G_i, Z_i) \geq \gamma_{G_i} \text{ and } X_i \in \mathcal{X}_{G_i}^P; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Academically efficient admissions: In the above setting, we define an admission practice to be academically efficient/fair if and only if γ_g is identical across g. The underlying intuition is that the only way covariates G should influence the admission process is through their effect on the expected academic outcome. 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_g will be lower for those gs which represent historically disadvantaged groups. Therefore, we are interested in identifying the value of the threshold γ_g for various values of g and testing if they are identical across g. We will call γ_g the "admission threshold" for group g.

It is important to note that here we are not making any assumption about whether or not G affects the distribution of the outcome, conditional on X. In our set-up, a male applicant with identical X as a female candidate can have a higher probability of being admitted and yet the admission process may be academically fair if males have a higher expected outcome than females with identical X. This contrasts sharply with the notion of fairness employed, for example, in Bertrand and Mullainathan (2004, BM) which concluded racial bias if two job applicants with identical CVs but of different race had different probabilities of being called for interview. In order for BM's finding to imply inefficiency according to our criterion, one needs to assume that, conditional on the information in CVs, race has no impact on average worker productivity.

4 Identification Analysis

In order to develop a test of meritocratic admissions, 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 $\varepsilon\,(\geq 0)$ 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 $SAT_i > SAT_j + \varepsilon\sigma_{SAT}$ and $GPA_i > GPA_j + \varepsilon\sigma_{GPA}$, where, σ_{GPA} and σ_{SAT} are the standard deviation of

[1] We assume that applicants with $x \notin \mathcal{X}_g^P$ are offered admission with probability 1 (if they are stronger than the best admitted candidate on whom data exist) or 0 (if they are worse than the worst admitted candidate on whom data exist).

GPA and SAT for the entire population of applicants.

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) $\mu_i^* = \phi\left(X_i, G_i, Z_i\right)$ (defined just before equation (3)) is continuously distributed conditionally on any realization of (X_i, G_i).

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

Assumption SD (Stochastic Dominance) There exists $\varepsilon > 0$ such that for any $e \geq \varepsilon$, if $x_g \in \mathcal{X}_g$ and $x_h \in \mathcal{X}_h$ with $x_g \succeq_e x_h$, then the distribution of Z conditional on $X = x_g$, $G = g$ first order 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) $\mu_i^* = \phi\left(X_i, G_i, Z_i\right)$ is continuously distributed conditionally on any realization of (X_i, G_i).

Discussion: Crudely speaking, Assumption M/SD means that applicants who are ε (or more) standard deviations better along standard, observable indicators of academic ability are likely to be viewed as 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 in our admission scenario, the outcome of interest Y is a measure of future academic performance in college 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 all past assessments are statistically more likely to have performed better in those assessments (unobserved by the researcher) which admission tutors view as positive determinants of college performance and hence, under the assumption of being academically motivated, would weigh positively in the decision to admit.

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 larger value of ε corresponds to a

weaker assumption, since $x_g \succeq_\varepsilon x_h$ will imply that the g-type individual is much better than the h-type one in terms of observables and hence it is more likely that the conclusion of Assumption M holds. In Subsection 7.2, below, we discuss a prescriptive method of choosing ε in generic applications, based on observables. We note however that for any choice of $\varepsilon \geq 0$, no matter how small, our identification relevant information (see section 4) will come from *all* pairs (g, h) where $x_g \succeq_\varepsilon x_h$, including those where the g-type is much better than the h-type in terms of observables.

Assumption M is similar in spirit but 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. Such characteristics may also be difficult to record for *past applicants* in a manner that is accessible by current admission tutors. However, while making admission 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. See Section 6 below for evidence that is suggestive and supportive of this assumption, for our application. 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 1 *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 2 *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.

In addition to Assumption M or SD, we will make a further assumption regarding the structure of $\phi(X_i, G_i, Z_i)$, viz.,

Assumption AS The tutors' subjective assessment ϕ satisfies

$$\mu_i^* \equiv \phi(X_i, G_i, Z_i) = \mu^P(X_i, G_i) + Z_i,$$

where $\mu^P(x, g)$ is defined in (2).[2, 3]

Discussion: Assumption AS concerns the structure of the "production" function $\phi(\cdot, \cdot, \cdot)$, as perceived by admission tutors, when faced with both "hard" information which is easy to record for past and current applicants and "soft information", observable to admission tutors only for the current applicants but otherwise difficult to record and hence unobservable to researchers. For example, tutors can infer the intellectual enthusiasm of each applicant in the current pool from his/her personal essay. But it is unlikely that tutors would remember such information about past cohorts, especially when faced with hundreds of applications to process every year. Therefore, a plausible method of selection is that when considering a current applicant, tutors form an initial impression of his/her future success – $\mu^P(X, G)$, based on the easily observable "hard" information like aptitude test score (e.g., SAT), high-school GPA etc. Then they adjust this initial impression, using an index of ability Z inferred from the "soft" information for each applicant in the current year which are unobserved by analysts (e.g., quality of reference letters and personal statements) to form the overall expectation $\mu^P(X_i, G_i) + Z_i$.

[2] Note that in general $\mu^P(x, g)$, will differ from $E[Y^P|X^P = x, G^P = g]$ which is typically unknown to admission tutors in universities because they, like us, do not observe potential outcomes of applicants who were not admitted or chose not to enrol. Indeed, a large literature in educational statistics on so-called "validation studies" use predicted performance of *admitted* candidates to infer the relative predictive ability of standardized test scores vis-a-vis high school grades and socioeconomic indicators and prescribe policies based on this analysis. See for example, Kobrin et al. (2001), Kuncel et al. (2008) and Sawyer (1996, 2010). Since our analysis evaluates what admission tutors are likely to do – rather than what one could have done under ideal circumstances like having experimental data – using $\mu^P(x, g)$ rather than $E[Y^P|X^P = x, G^P = g]$ – is the correct approach here. Obviously, under selection on observables, these two quantities are identical.

[3] We are implicitly assuming that regressing outcome data for past applicants observed by the analyst yields a consistent estimate of $\mu^P(X, G)$ used by admission-tutors, which is likely when tutors rely on more recent data, rather than historical data unobserved by analysts, to make predictions.

Assumptions AS and M yield a lower bound on the threshold differences. To see this, define the function

$$p(x,g) := \Pr[D=1|X=x,G=g],$$

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

$$\mathcal{M}(g,h,\varepsilon) := \{(x_g,x_h) \in \mathcal{X}_g \times \mathcal{X}_h : x_g \succeq_\varepsilon x_h,\ p(x_g,g) \leq 0.5 < p(x_h,h)\}. \tag{4}$$

Note that the set $\mathcal{M}(g,h,\varepsilon)$ can be directly computed from the data because it depends only on observables. Also, let $Q^\alpha[Z|x,g]$ denote the αth quantile of the random variable Z, conditional on $(X,G)=(x,g)$, with $\alpha = 0.5$ corresponding to the median. Now, note that

$$\begin{aligned} 1-p(X_g,g) &: = 1-\Pr[D=1|X=x_g,G=g] \\ &= \Pr\left[Z < \gamma_g - \mu^P(x_g,g)\,|X=x_g,G=g\right]. \end{aligned}$$

This implies that

$$\gamma_g = \mu^P(x_g,g) + Q^{1-p(x_g,g)}[Z|x_g,g],$$

since Z is continuously distributed (by part (ii) of Assumption M). Similarly for individuals with $(X,G)=(x_h,h)$ with $g \neq h$,

$$\gamma_h = \mu^P(x_h,h) + Q^{1-p(x_h,h)}[Z|x_h,h].$$

Then,

$$\gamma_g - \gamma_h = \mu^P(x_g,g) - \mu^P(x_h,h) + Q^{1-p(x_g,g)}[Z|x_g,g] - Q^{1-p(x_h,h)}[Z|x_h,h].$$

Now if $p(x_g,g) < 0.5 \leq p(x_h,h)$, then

$$\begin{aligned} \gamma_g - \gamma_h &> \mu^P(x_g,g) - \mu^P(x_h,h) + Q^{1-0.5}[Z|x_g,g] - Q^{1-0.5}[Z|x_h,h] \\ &= \mu^P(x_g,g) - \mu^P(x_h,h) + \text{Median}[Z|x_g,g] - \text{Median}[Z|x_h,h]. \end{aligned}$$

So if in addition, $x_g \succeq_\varepsilon x_h$, then by Assumption M, $\text{Median}[Z|x_g,g] \geq \text{Median}[Z|x_h,h]$ and hence

$$\gamma_g - \gamma_h > \mu^P(x_g,g) - \mu^P(x_h,h).$$

Taking the supremum of the RHS over (x_g,x_h) satisfying $(x_g,x_h) \in \mathcal{S}(g,h,\varepsilon)$ and $(x_g,x_h) \in \mathcal{X}_g^P \times \mathcal{X}_h^P$ (so that we can compute $\mu^P(x_g,g) - \mu^P(x_h,h)$ for all these pairs), we get

$$\gamma_g - \gamma_h \geq \sup_{(x_g,x_h)\in\mathcal{M}(g,h,\varepsilon)} \left[\mu^P(x_g,g) - \mu^P(x_h,h)\right] \equiv \underline{\theta}(g,h). \tag{5}$$

The RHS of the above inequality is based only on observables and is easy to compute once we specify regression models for $\mu^P(\cdot,\cdot)$ and $p(\cdot,\cdot)$. If this lower bound is positive, then we can conclude that group g is facing a higher admission threshold. Under the stronger condition of Assumption SD, we can analogously define

$$\mathcal{SD}(g,h,\varepsilon) := \{(x_g, x_h) \in \mathcal{X}_g \times \mathcal{X}_h,\ x_g \succeq_\varepsilon x_h,\ p(x_g, g) \leq p(x_h, h)\}, \tag{6}$$

whence we have the bound

$$\gamma_g - \gamma_h \geq \sup_{(x_g,x_h)\in\mathcal{SD}(g,h,\varepsilon)} \left[\mu^P(x_g, g) - \mu^P(x_h, h)\right]. \tag{7}$$

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. Therefore, if these $g$type applicants have higher predicted performance based on observables, then they must have been facing a higher threshold. Other pairs of applicants for whom dominance does not hold do not contribute to the identification.

4.1 Alternative identification strategies

We are not aware of any existing empirical method of identifying the extent of affirmative action or of rigorously testing *outcome-based* efficiency of college admissions. 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, which is known to be a key difficulty in detecting who the marginal treatment recipients are. For example, in the admission context, it is quite likely that students from disadvantaged backgrounds have larger mean and variance in academic ability, conditional on having obtained the same score in school-leaving examinations as students from wealthier backgrounds. Our analysis imposes no such restriction on the unobservables' distribution. In the healthcare context, Bhattacharya (2013) suggests an alternative approach to testing outcome-oriented treatment assignment via a partial identification analysis using a combination of observational data and prior experimental findings from randomized controlled trials. Such experimental results are typically difficult to come by in the college admission context.

In other contexts such as law-enforcement and healthcare provision, researchers have used economic optimization based reasoning to detect racial prejudice (c.f. Persico, 2009 for a survey). For instance, Knowles, Persico and Todd (2004) evoke the assumption that potential criminals respond optimally to drug-enforcement protocols by adjusting the amount of contraband they carry. This insight justifies the equating of the *unobservable marginal* with the *observable average* outcomes across "treated" individuals (i.e., motorists who are apprehended) and thus can be used to test whether marginal outcomes are equated across demographic groups. 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 sample-based 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 part B of the Appendix) is strong and conducting statistical inference under it is slightly simpler. The first task regarding inference is to test whether $\mathcal{SD}(g,h,\varepsilon)$ (defined in (6)) is nonempty. Indeed, $\underline{\theta}(g,h)$ is well-defined only when $\mathcal{SD}(g,h,\varepsilon)$ is nonempty while it is $-\infty$ otherwise, suggesting that we have no information about $\gamma_g - \gamma_h$. We will focus on the case where $\mu(\cdot,\cdot)$ and $p(\cdot,\cdot)$ are parametrically specified via linear and probit models, respectively. That is,

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

where $(\boldsymbol{\beta}_{0,g},\boldsymbol{\beta}_{0,h})$ and $(\boldsymbol{\delta}_{0,g},\boldsymbol{\delta}_{0,h})$ are the true linear-regression and probit coefficients; and Φ is the C.D.F. of the standard normal. In principle, one can also use nonparametric estimates for $\mu(\cdot,\cdot)$ and $p(\cdot,\cdot)$ but due to relatively small sample-size, the two-sample nature of the problem and the complicated construction of "intersection bounds" for nonparametric estimates (needed for testing emptiness), we do not consider such methods here. 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

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

Testing emptiness: Observe that the null hypothesis of an empty $\mathcal{SD}(g,h,\varepsilon)$ is equivalent to the hypothesis that $\theta_0 \geq 0$, where

$$\theta_0 := \inf_{(x_g,x_h)\in\mathcal{X}_g\times\mathcal{X}_h,\ x_g\succeq_\varepsilon x_h} [p(x_g,g) - p(x_h,h)]$$

The quantity θ_0 is of a form analyzed in Chernozhukov et al (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.

Quantile-based lower bound estimator and its asymptotic distribution: For the application, when bounding the magnitude of threshold differences, we consider a slightly more conservative bound which is easier to conduct inference on. Note from (5) that the key parameter of interest is a supremum over the domain $\mathcal{SD}(g,h,\varepsilon)$, defined in (6). Now, since $p(x_g,g)$ needs to be estimated, we need to conduct inference on the supremum of an estimated object, viz., $\mu^P(x_g,g) - \mu^P(x_h,h)$ *over an estimated domain.* This problem is not covered by existing methods in the literature on partial identification or moment inequalities. Instead of developing distribution theory for this supremum, we will work with a slightly conservative version of the bound, viz., we replace the supremum $\underline{\theta}(g,h)$ (defined in (5)) by the upper λth quantile, and conduct inference on it. That is, we use the *implication* of (5) that for any $\lambda \in (0,1)$,

$$\gamma_g - \gamma_h \geq \underline{\theta}(g,h) \geq \theta_0^\lambda(g,h), \tag{8}$$

where $\theta_0^\lambda(g,h)$ is the λth quantile of the difference in (5):

$$\theta^\lambda(g,h) := Q^\lambda\left[\mu^P(X_g,g) - \mu^P(X_h,h) \,\middle|\, \begin{array}{c}(X_g,X_h)\in\mathcal{X}_g\times\mathcal{X}_h, X_g\succeq_\varepsilon X_h,\\ p(X_g,g)\leq p(X_h,h)\end{array}\right]. \tag{9}$$

For any λ (bounded away from 0 and 1), we obtain a corresponding lower bound for $\gamma_g - \gamma_h$. If $\theta_0^\lambda(g,h)$ is larger than zero, then so is $\underline{\theta}(g,h)$ and thus we can conclude that $\gamma_g > \gamma_h$. In the application, we show results for $\lambda = 0.80$. In the terminology of partial identification analysis, this is analogous to calculating an "outer identification region" for model parameters. Our estimator of $\theta_0^\lambda(g,h)$ is the natural sample analog of (9):

$$\hat{\theta}^\lambda(g,h) = \hat{Q}^\lambda\left[\hat{\mu}^P(X_g,g) - \hat{\mu}^P(X_h,h)\,|X_g\succeq_\varepsilon X_h,\ \hat{p}(X_g,g)\leq\hat{p}(X_h,h)\right],$$

where X_g is associated with $G=g$, and X_h with $G=h$; $\hat{Q}^\lambda$ is the λth quantile based on the empirical distribution of (X_g,X_h); and $\hat{\mu}^P$ and $\hat{p}$ are functions estimated in a preliminary step.

This can be stated as two-sample moment condition problem where the moments are nonsmooth in the parameters. As such, the distribution theory for obtaining confidence intervals for $\theta_0^{\lambda}(g,h)$ does not follow directly from existing results in the econometrics literature and requires an independent analysis. In an online appendix posted on the second author's website, we show that the asymptotic distribution of the eventual estimator $\hat{\theta}^{\lambda}(g,h)$ is asymptotically normal with a consistently estimable asymptotic variance. Based on the estimate of the asymptotic variance, we can construct confidence intervals for the lower bound $\theta_0^{\lambda}(g,h)$.

6 Empirical Analysis

Background: Our empirical analysis is based on admissions data for three recent cohorts of applicants to an undergraduate degree programme in a popular subject at a selective UK University. Like in many other European and Asian countries, students enter British universities to study a specific subject from the start, rather than the US model of following a broad general curriculum in the beginning, 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 other applicants 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, such as leadership qualities, suitability as team-members, engagement with the community etc., 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 – e.g., Math, History, English, Physical and Biological Sciences etc. The examinations are centrally conducted and hence scores of individual students on these examinations are directly comparable, unlike high-school GPA in the US where candidates undergo school-specific assessments which may not be directly comparable across schools. 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 focus on UK-based 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 aptitude-test and essay scores are separately recorded. All of

this information is then entered into a spread-sheet held at a central database which all admission tutors can access. About one-third of all applicants are selected for interview by the university on the basis of UCAS information, aptitude test and essay, and the rest rejected. Selected candidates are then assessed via a face-to-face interview and the interview scores are recorded in the central database. 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 the above information from among the candidates called for interviews. For our application, we use anonymized data for three cohorts of applicants from their records held at the central admissions database at the university. For the admitted students, we merged these with their performance in the first year, in which students sit tests in three papers. The scores across the three papers are averaged to calculate the overall performance, which we take to be the outcome of interest. As an alternative and for comparison, we consider performance in the final examinations taken at the end of three years in eight papers. We have the finals data for the two earlier cohorts and not the third one.

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 X 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 is thus based on recommendation letters, the applicant's personal essay (not the substantive essay they write as part of the aptitude test), any prizes or distinctions obtained among possibly other indicators. Given that those summoned for interview constitute our "population" of interest, we found that in terms of A-level grades, GCSE scores and 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 (see the Introduction). 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. 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 appear to have lagged behind the trend: in 2010-11, 55% of undergraduates across all UK universities were female, but 44% of students admitted to the 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.

Outcome: After entering university, the candidates take preliminary examinations in three papers at the end of their first year. Each script is marked blindly, i.e., the marking tutors do not know anything about the candidate's background or gender. We use the average score over the three papers as the first outcome – labelled prelim_tot – which can range from 0 to 100. An advantage of using the preliminary year score as the relevant outcome measure is that every admitted student sits the same preliminary exam in any given year; so there is no confounding from the difference in score distributions across different optional subjects, as often happens in the final examinations at the end of the 3-year course. The disadvantage of using the first year score is that applicants from relatively modest socioeconomic backgrounds are more likely to "catch up" at the end of three years and thus an assessment based on prelim scores may bias a researcher towards overestimating the extent of affirmative action.

In view of these considerations, we use as a second outcome the students' performance in the final examinations in eight papers which are taken at the end of three years and based on which the student receives his/her degree. At this stage, students do not all sit the same papers; but the marking is still blind and the scores reflect relative competence with respect to the others taking the same paper. The disadvantage of this outcome is that students take examinations in different papers which they self-select into and therefore any real improvement relative to the first-year is, to some extent, confounded with efficient sorting into options. Using Duke University data, Arcidiacono et al. (2011) have recently documented large differences in patterns of major choice between candidates who are the likely beneficiaries of affirmative action policies during admissions compared to the major choice patterns of other enrolled students. However, unlike in Arcidiacono et al., here the sorting is not into easier and harder subjects (like STEM and non-STEM majors) but only into different options which are intellectually similarly demanding.

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 and male admits score an average of about 1 percentage point (20% of the overall standard deviation) higher in the first year exams. 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 gcsescore.

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).

Evidence of median-dominance: Among the pre-admission variables that we observe in our dataset it's only the performance in the interview that 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 by treating the interview score as the unobservable component. That is, we will verify whether the median interview score is stochastically higher for those types of applicants who are better in terms of all other "tutor-independent" test-scores obtained in prior assessments. If that is true, then our Assumption M regarding the truly unobservable determinants of admissions is also more credible. The concrete steps leading to our test are as follows. Consider $X = (Aptitude\_test\_score, Exam\_essay)'$. 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 2, 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 2 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 hold for all higher values of ε.

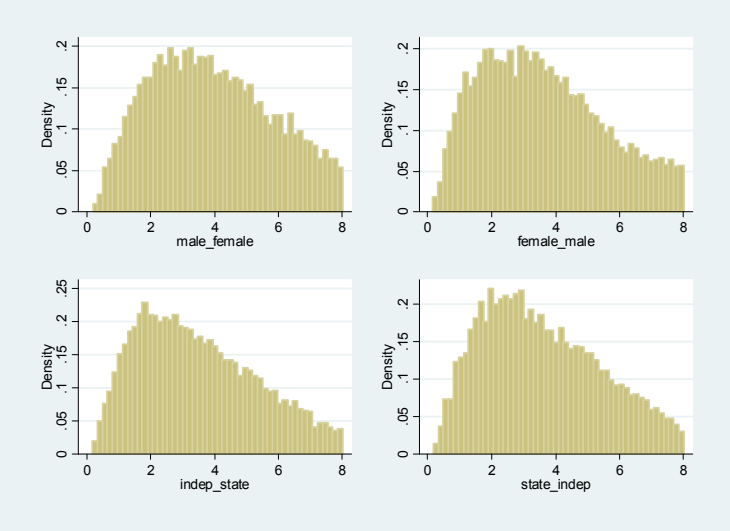


Figure 2: Evidence of Median Dominance

It is evident that all four of these histograms have entirely positive support, suggesting that the median dominance conditions hold. 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.

6.1 A thought experiment

Before performing empirical analysis of the actual data, we conduct a thought experiment where we investigate the usefulness of our approach in a situation where the "truth" is known. The idea is to treat one of the observed covariates – viz., the interview score – as unobserved, note that this "missing" covariate satisfies our assumption of median monotonicity (see Figure 2) and then run a simulation experiment where tutors accept applicants based on all characteristics including the interview score but the researcher does not observe it. In this simulation experiment, we vary the acceptance thresholds and check how small a difference in thresholds can our bounds-based method detect when the interview score remains "unobserved" to us.

Simulation exercise: We now conduct a simulation exercise, whose purpose is to investigate how well our method works when we a priori know the admission thresholds. In order to do this, we use the above dataset where we treat a school type as G, and aptitude test and examination essay scores and gender as the commonly observed covariates, X. The interview score is taken to be unobserved by us (researchers) but observed by admission tutors for the present cohort for whom the admission decision is to be made. This will play the role of Z. We generate artificial observations on admissions in the following way. Using past academic performance in the first year

examination as the outcome, we estimate a regression model where X are used as regressors. We then generate the predicted outcomes for each current year applicant by using coefficient estimates from the previous regression and adding a contribution from the "unobserved" interview score Z_1 (normalized to have mean zero across the entire sample). If this sum plus a stochastic slippage error exceeds a threshold value of 61.5 for state-school students ($G = h$) and $61.5 + \delta$ for independent school applicants ($G = g$), then the student is assumed to have been offered admission, i.e., the admission-dummy D is set to be 1. It is set to be 0 otherwise. That is, we set

$$\begin{aligned}
\boldsymbol{\beta}_g &= \left[\sum_{i=1}^{n} \mathbf{1}\{G_i = g\} X_i X_i'\right]^{-1} \sum_{i=1}^{n} \mathbf{1}\{G_i = g\} X_i Y_i; \\
\boldsymbol{\beta}_h &= \left[\sum_{i=1}^{n} \mathbf{1}\{G_i = h\} X_i X_i'\right]^{-1} \sum_{i=1}^{n} \mathbf{1}\{G_i = h\} X_i Y_i; \\
D_i &= \mathbf{1}\left\{X_i'\boldsymbol{\beta}_g \mathbf{1}\{G_i = g\} + X_i'\boldsymbol{\beta}_h \mathbf{1}\{G_i = h\} + 0.05 Z_{1,i} + u_i \geq 61.5 + \delta \times \mathbf{1}\{G_i = g\}\right\},
\end{aligned}$$

where $0.05Z_{1,i}$ is the contribution from an "unobserved" interview score; u_i is the stochastic slippage component drawn from the normal distribution $N(0, \mathbf{1}\{G_i = g\} + 2 \times \mathbf{1}\{G_i = h\})$ and thus the sum $0.05Z_{1,i} + u_i$ represents the unobserved index variable Z_i; and finally, δ, which is set externally by us, is the extent of affirmative action. A positive value of δ indicates that independent school applicants are being held to a higher threshold of expected performance.

For each value of δ, we then perform our bounds analysis by pretending that we observe X but not the interview score. This is meant to capture the situation that admission tutors may base their decision on some subjectively assessed performances Z, unobserved by the researcher, in addition to the prediction based on the commonly observed covariates. Since the interview-score satisfies Assumption M (see Figure. 2), our bounds analysis is applicable in this case. Accordingly, Table 3 reports true values of δ and the corresponding lower bounds on it, obtained by using our method with $\lambda = 0.5$ (median), $\lambda = 0.80$ as well as the mean. The table can be read as follows. The first column reports the true value of δ, the second column shows the fraction of times we have $p(x_g, g) < p(x_h, h)$ among all pairs satisfying $x_g \succeq_\varepsilon x_h$ with $\varepsilon = 0.1$. The point estimates for median, mean and 80th percentile of the difference $\mu(X_g, g) - \mu(X_h, h)$ over $\mathcal{SD}(g, h, \varepsilon)$ are reported in the next three columns. Finally, equal tailed confidence intervals (obtained by repeated sampling from this design) are reported below the estimates.

It can be seen from Table 3 that threshold differences of 2 or more points out of 100 (overall standard deviation of the outcome distribution is about 5 points) are clearly detected; a positive difference of 1 or less still yields positive point-estimates for δ but the associated confidence intervals contain 0. For a negative value of δ, the fraction f is calculated to be zero, as one would expect. Overall, this table presents strong evidence that our method works well in practice. As such, this

exercise increases the credibility of the results obtained by applying our methods to the full dataset where interview scores are observed together with the covariates included in X but some other characteristics which are potentially used by tutors in predicting future performance, may remain unobserved to us.

7 Results

We now turn to the real application where we use the aptitude test score, the examination essay 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 will 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 $\mathcal{SD}(g, h, \varepsilon)$. We first do this graphically by plotting the marginal C.D.F. of the difference in acceptance 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.[4] 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.

[4] 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 $S_\varepsilon^{SD}(g, h)$.

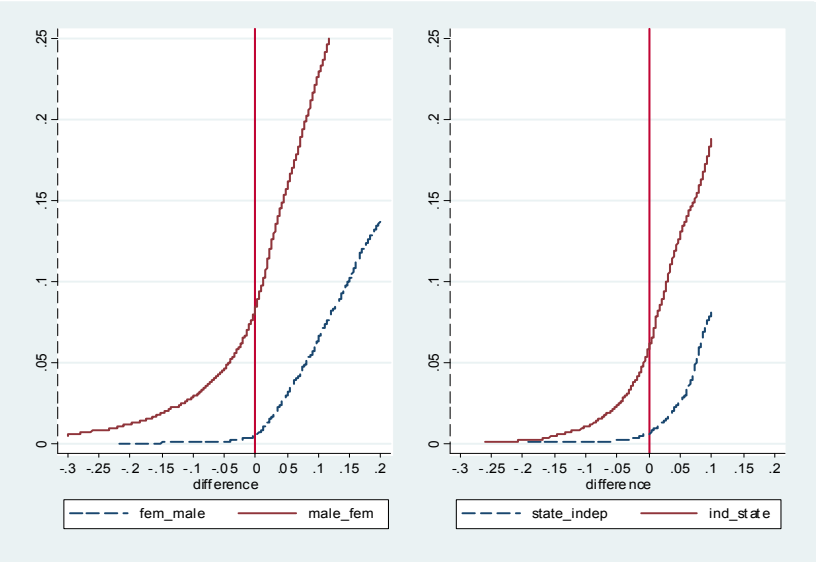


Figure 3: Evidence of Emptiness

Clearly, the first curve has significant mass below zero and the dashed curve has almost no mass below zero, suggesting a positive probability that $p\left(X_{male}, male\right) < p\left(X_{female}, female\right)$ 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 4, we report $\hat{\theta}_{0n}\left(0.95\right)$, the upper limit of a one-sided confidence interval, calculated using the method of CLR, as explained in Section 5. A negative upper limit indicates that the set $\mathcal{SD}\left(g, h, \varepsilon\right)$ 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 Table 4 that we reject emptiness for $g = male$, $h = female$ and for $g = indep$, $h = state$ but do not reject emptiness in the other cases.

Given the conclusion of the test of emptiness, we now compute lower bounds for $\gamma_{male} - \gamma_{female}$ and $\gamma_{indep} - \gamma_{state}$, based on $\lambda = 0.8$ (c.f. eq. (9)). We use a value of $\varepsilon = 0.1$ and later we compare estimates obtained using $\varepsilon = 0.25$ with those obtained using $\varepsilon = 0.1$. In Tables 5A and 5B we report the estimated lower bounds $\hat{\theta}^{\lambda}$ for $\lambda = 0.80$, given by (5) and (7), using prelim and finals performance as outcomes, respectively. The first column, labeled "upper limit", reports $\hat{\theta}_{0n}\left(0.95\right)$ from the previous table. When this number is negative, it indicates that the $\mathcal{SD}\left(g, h, \varepsilon\right)$ is nonempty, whence we proceed to compute $\hat{\theta}^{\lambda}$. Imposing the assumption AS and calculating lower bounds on the magnitude of the threshold differences, we get values of 3.78 and 2.14 for gender and school-type, respectively, suggesting that the marginal male admits and the marginal independent school admit perform significantly better in their first year examinations. In terms of the overall

distribution of first year exam scores, these differences amount to about 65% and 40%, respectively, of one standard deviation.

Comparing these results with the finals performance reported in Table 5B, we see that the magnitude of the lower bound has now shrunk by more than 50%. That is, the marginal male admit is expected to perform at least 1.95 points higher than the marginal female admit. This gender difference is still significant but the one for school-type is not. Since it is the lower bound which has shrunk, it is not immediate whether the actual difference has also shrunk. However, the large magnitude difference does suggest some shrinking of the actual gaps resulting from either catch-up over time and/or some extent of efficient sorting into options.

Table 5C reports the bounds where the outcome is the indicator of whether a student gets a first-class mark (i.e. 70% or more) in the finals. Eventual degrees of all graduating students are classified into four categories – first, upper second, lower second and pass. On average, approximately 25-30% of students get a first-class degree in the subject we are studying; a first-class degree from university is associated with significant academic prestige and significantly improves one's chances of admission to high-ranked post-graduate programmes and prospects of securing attractive jobs beyond graduation. It can be seen from Table 5C that the qualitative conclusions remain the same as before – a large gender gap exists but the gap across school-types is insignificant. Finally, in Table 6, we compare estimates using $\varepsilon = 0.25$ with those obtained using $\varepsilon = 0.1$. The differences in results can be seen to be very small.

The exact magnitudes of the lower bounds reported in Tables 5-6 vary slightly across functional specifications (e.g. whether higher order terms and interactions in the test scores are or are not used to estimate $\mu^P(\cdot,\cdot)$), but three empirical findings are robust across all specifications: (a) the gender gap is large, persistent and statistically significant in every case; (b) the independent-state school difference is comparatively smaller; and (c) the lower bounds based on the final-year examinations are smaller then the ones based on first-year performance but the gender gap in admission thresholds remains significant.

In order to gain some visual insight into how the threshold discrepancies arise, in Figure 4, we plot the empirical marginal C.D.F.'s of the estimated $\mu^P(X_{male}, male)$ and $\mu^P(X_{female}, female)$ (the left panel) and those of the estimated $\mu^P(X_{indep}, indep)$ and $\mu^P(X_{state}, state)$ (the right panel). Here we take first-year performance as the outcome of interest. It is clear that the male distribution first-order stochastically dominates the female distribution. This means that if admissions are deterministic, conditional on μ^P (i.e., there is no unobserved heterogeneity), *any* common acceptance rate across gender will result in a higher μ^P for the marginal accepted male than the marginal

accepted female.

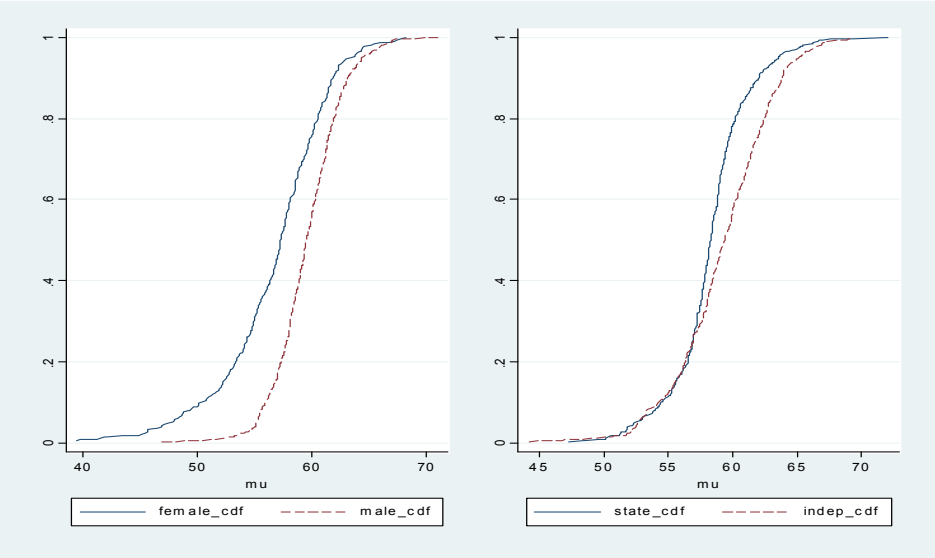


Figure 4: Graphical Illustration of higher threshold

This can be seen in Figure 4, by looking along any fixed cutoff on the vertical axis. Any such horizontal cut-off line[5] will intersect the female C.D.F. at a point that will lie strictly to the left of the point of intersection with the male C.D.F. Given the results presented in the tables, it is evident that the presence of unobserved heterogeneity does not alter this fundamental dominance situation. A similar, albeit relatively weaker, dominance situation occurs for school-type, as can be seen in the right-hand panel in Figure 4.

Interpretation of the empirical findings: It would be natural to conjecture that the observed 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 1A and 1B). This would make the effective male threshold higher if, say, the conditional male outcome distribution has a thicker right tail (see Figure 4). 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 1A 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 or independent school applicants face effectively

[5] For instance, if the top 30% of applicants are accepted among both males and among females, then we should be looking along the horizontal line at $1 - 0.3 = 0.7$ on the vertical axis.

higher admission thresholds. Some of the difference observed in the first-year performance is mitigated when one examines the final-year performance. This is likely caused by a combination of academic improvement by the lower performing students who are female or from state-schools and a more efficient sorting into optional courses. As explained above, the various optional courses do not necessarily pose different levels of intellectual challenges and hence the observed shrinking of the initial differences should be viewed as a sign of relative improvement by those performing the worst in the first-year exams.

7.1 Robustness of interpretation

We now investigate whether our findings could be consistent with two alternative explanations.

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; the question is whether by including G in our analysis, we are "detecting" threshold differences that are not there in the actual admission process. Even if this is the case, we would argue that in order for admissions to be meritocratic, admission tutors should take G into account. For example, suppose G denotes a school type, state-school students are more able than independent school students with the same test score, and therefore perform better in university 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. Nonetheless, for interpreting our finding of different thresholds, one might investigate G-blindness as a possible explanation. Accordingly, let $\bar{\mu}^P(X)$ denote the expected future performance based on X but not G and consider an alternative admission rule

$$D = \mathbf{1}\left\{\bar{\mu}^P(X) + Z \geq \gamma_G\right\},$$

where, under a G-blind admission process, γ_G will not vary by G. Now, for $x_g \in \mathcal{X}_g$,

$$p(x_g, g) := \Pr[D = 1 | (X, G) = (x_g, g)] = \Pr[Z \geq \gamma_g - \bar{\mu}^P(X) | (X, G) = (x_g, g)],$$

Then, we have

$$\gamma_g = \bar{\mu}^P(x_g) + Q^{1-p(x_g,g)}[Z|x_g, g].$$

Similarly, for $x_h \in \mathcal{X}_h$,

$$\gamma_h = \bar{\mu}^P(x_h) + Q^{1-p(x_h,h)}[Z|x_h, h],$$

and thus

$$\gamma_g - \gamma_h = \bar{\mu}^P(x_g) - \bar{\mu}^P(x_h) + Q^{1-p(x_g,g)}[Z|x_g, g] - Q^{1-p(x_h,h)}[Z|x_h, h],$$

implying, under Assumption M, that

$$\gamma_g - \gamma_h \geq \sup_{(x_g, x_h) \in \mathcal{SD}(g,h,\varepsilon)} \left[\bar{\mu}^P (x_g) - \bar{\mu}^P (x_h) \right],$$

where $\mathcal{SD}(g, h, \varepsilon)$ is defined exactly as above. If the supremum exceeds zero, then we can conclude that admissions were not generated in a fully G-blind way. The RHS lower bound is similar to (5) except that $\mu^P(\cdot)$ is not conditioned on G. We compute the 80th percentile instead of the supremum, as before and report this in column 1 of the following table (under the heading "G-blind"), for $\varepsilon = 0.1$ and for the outcome being the finals performance.

Alternative Interpretations

Category	G-blind	No-Interview	Benchmark
Male-Female	1.97	2.85	1.93
Indep-State	1.65	0.96	0.75

The table shows that the threshold differences are in fact slightly *larger* if we assume that G is not used to predict future outcomes and thus G-blind admissions are unlikely to be an explanation.

Biased interviews scores: 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 2, 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 2, which is clearly not the case. Furthermore, our method of identifying threshold differences is based on the predicted performance in university exams as a function of interview and other test-scores, rather than the test scores in themselves. Under biased interview scores, g-type candidates with low ability but high interview scores (due to the bias) will perform relatively poorly upon being admitted and thus have *lower* values of $\mu^P(x, g)$ for fixed x. This will make our bounds, based on the difference $\mu^P(x, g) - \mu^P(x_h, h)$ for those with $p(x, g) < p(x_h, h)$, negative (or less positive). So interpreting a *positive* lower bound as symptomatic of nonacademic bias against g-type candidates is robust to interview scores being biased in favor of g-type applicants. The bounds obtained upon ignoring interview scores altogether are reported in the third column of the previous table. The lower bound on the male-female difference is now much *larger* than the benchmark case and the independent-state difference

similar in magnitude (both being statistically insignificant). Thus our substantive conclusions remain valid.

8 Summary and Conclusion

This paper has proposed an empirical method for 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. Our approach works by obtaining bounds on the *difference* in admission thresholds faced by applicants of different demographic groups. These bounds are robust to the unobserved characteristics problem, under an intuitive assumption about the ranking of applicants by unobservable attributes. The bounds reveal information about the extent of bias in the admission process relative to the meritocratic ideal of admitting students with the highest academic potential. Since our methods are based on predicted probability of acceptance and predicted performance in university, they can be applied to situations where applicants come from diverse backgrounds and report scores from different aptitude tests, since the necessary predicted values can be calculated based on candidate-specific covariates. Furthermore, we do not require any information for past applicants who were not accepted, which is convenient since universities normally do not store such data. 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 slightly higher relative to state school applicants. In contrast, average admission rates are nearly identical across gender and across school-type, both before and after controlling for other covariates.

We have left several substantive issues to future research. For example, we do not consider peer-effects in our analysis because it is unlikely that admission tutors have enough information regarding peer effects to base their admission decisions on it. Second, it may be useful to repeat the empirical analysis using other outcomes – such as wage upon graduation – which are more directly related to social mobility. However, we suspect that college performance data are much more readily available than wage data because the latter requires costly follow-up of alumni and can entail non-ignorable non-response. Lastly, our methods are potentially useful for testing outcome-based fairness of binary decisions in non-admission contexts such as approval of mortgage applications, referrals to expensive medical treatment etc., where allegations of unfair decision are common and where eventual outcomes are observed for those who were approved or treated.

Table 0: Variable-Label

gcsescore	Overall score in GCSE, 0-4
alevelscore	Average A-level scores 80-120
took subject 1	Whether studied 1st recommended subject at A-level
took subject 2	Whether studied 2nd recommended subject at A-level
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
finals_avg	Average Score in final year examination; 0-100
offer	Whether offered admission
accept	Whether accepted admission offer

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. 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 University recommends that candidates study two specific subjects at A-levels for entry into the undergraduate programme under study. Subject 1 and Subject 2 are dummies for whether an applicant did study them at A-level.

Table 1. Means by Gender and by Schooltype

Variable	**Female (N=365)**	**Male (N=620)**	**pvalue_diff**	**State (N=548)**	**Indep (N=437)**	**pvalue_diff**
gcsescore	3.83	3.75	0	3.70	3.87	0
took subject 1	0.69	0.68	0.54	0.64	0.73	0.02
took subject 2	0.48	0.52	0.27	0.53	0.49	0.004
alevelscore	119.73	119.44	0.01	119.60	119.73	0.02
aptitude test	62.53	65.24	0	63.82	64.94	0.0015
essay	63.23	64.49	0	64.06	64.07	0.5
interview	64.23	65.29	0.04	65.02	65.17	0.65
Prelim_avg	60.98	61.89	0.04	61.15	62.10	0.03
Finals_avg	64.89	65.34	0.28	65.02	65.37	0.88
offer	0.363	0.357	0.41	0.361	0.357	0.5
accept	0.34	0.34	0.5	0.33	0.35	0.46

Note: The data pertain to three cohorts of applicants. The variable names are explained in table 0. Column 6 records 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%.

Table 2. Probit of receiving offer

Regressor	Coef.	Std. Err.	z	p-value
gcsescore	0.26	0.25	1.04	0.30
alevelscore	0.08	0.06	1.26	0.21
took subject 1	-0.06	0.17	-0.33	0.74
took subject 2	-0.25	0.15	-1.65	0.10
aptitude test	0.09	0.01	7.01	0.00
essay	0.01	0.01	0.44	0.66
interview	0.23	0.02	10.59	0.00
indep	-0.13	0.15	-0.88	0.38
male	-0.18	0.16	-1.13	0.26

N=985, Pseudo-R-squared=0.5

Table 3: Simulation: Indep-State

True difference, δ	Fraction_negative	Median	Mean	80%ile
4	24.2	3.32	3.35	4.02
		(1.35, 4.93)	(1.59, 5.02)	(2.21, 6.21)
3	17.7	1.92	1.88	2.63
		(0.28, 3.22)	(0.05, 3.01)	(1.96, 4.07)
2	12.99	1.33	1.31	1.54
		(0.76, 2.78)	(0.28, 2.31)	(0.51, 2.88)
1	3.12	0.86	0.86	1.21
		(0.14, 1.60)	(-0.08, 1.36)	(-0.88, 1.99)
0	0.2	-0.06	0.11	0.29
		(-1.88, 0.47)	(-1.78, 0.47)	(-1.45, 0.61)
-2	0	.	.	.

Note: Results of Simulation exercise as described in section 6.1 of text. The first column is the true threshold difference used in the simulation. Column 2 reports the fraction of covariate pairs satisfying the relation described by the set S(g,h) among all possible covariate pairs. A larger fraction indicates that the set S is more likely to be non-empty. The last thress columns report the estimated lower bounds on the threshold differences, based on the median, mean and 80th percentiles of the conditional mean differences over the set S(g,h).

Table 4: Test Emptiness of S(g,h) for ε=0.1

Difference	Upper limit of CLR CI
g=male, h=female	-1.53
g=female, h=male	0.35
g=indep, h=state	-0.33
g=state, h=indep	0.79

Upper limit of 95% confidence interval for a test of empty conditioning set S(g,h) based on CLR.Negative value indicates non-empty set and implies that group g faces a higher threshold, resulting from assumptions CM and SD in the text.

Table 5A: Threshold Differences, Prelim, Mean=61.58, s.d.=5.91

Difference	test emptiness	lower bd: 80 %ile	pvalue lower bd
male-female	-1.53	3.78	0.07
female-male	0.35	.	.
indep-state	-0.33	2.14	0.04
state-indep	0.79	.	.

Table 5B: Threshold Differences, Finals, Mean=64.94, s.d.=4.22

Difference	test emptiness	lower bd: 80 %ile	pvalue lower bd
male-female	-1.53	1.95	0.09
female-male	0.35	.	.
indep-state	-0.33	0.75	0.46
state-indep	0.79	.	.

Table 5C: Threshold difference, Outcome=First Class, Mean=0.30

Difference	test emptiness	lower bd: 80 %ile	pvalue lower bd
male-female	-1.53	0.153	0.02
indep-state	-0.33	0.094	0.25

Notes: The data pertain to three cohorts of applicants. The first column presents upper limit of 95% confidence interval for a test of empty conditioning set based on CLR; negative value indicates non-empty set. Upon rejecting emptiness, we compute lower bound on threshold differences and corresponding p-values, as per subsections 4.4 and 5.2 of text.

Table 6: Threshold Differences for different ε

PRELIM, Mean=61.58, s.d.=5.91			FINALS, Mean=64.94, s.d.=4.22		
ε	**0.1**	**0.25**	**ε**	**0.1**	**0.25**
male-female	3.78	3.11	male-female	1.95	2.03
pvalue	0.07	0.1	pvalue	0.09	0.12
indep-state	2.14	2.64	indep-state	0.75	0.99
pvalue	0.04	0.04	pvalue	0.46	0.5

References

[1] Altonji, J.G. & Blank, R.M. (1999) Race and gender in the labor market, Handbook of Labor Economics Vol. 3C (O. Ashenfelter and D.Card, eds.), 3143-259. Elsevier, New York.

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

[3] Arcidiacono, P., E. M. Aucejo & K. Spenner (2011) What happens after enrollment? An analysis of the time path of racial differences in GPA and major choice?, working paper, Duke University.

[4] Becker, G. (1957) The economics of discrimination, University of Chicago Press.

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

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

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

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

[9] Blank, R., M. Dabady & C. Citro (2004): Measuring Racial Discrimination, Washington, D.C.: National Research Council, National Academy Press.

[10] Card, D. & A.B. 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.

[11] Chandra, A. & D. Staiger (2009) Identifying provider prejudice in medical care, Mimeo, Harvard University and Dartmouth College.

[12] Chernozhukov, V., S. Lee & A. Rosen (2013) Intersection bounds: Estimation and inference, Econometrica, 81-2, 667-737.

[13] Charles, K. & J. Guryan (2011) Studying discrimination: Fundamental challenges and recent progress. Annual Review of Economics, 3, 479–511.

[14] Guryan, J. & K. Charles (2013) Taste-based or statistical discrimination: The economics of discrimination returns to its roots, forthcoming in Economic Journal.

[15] Fryer Jr., R.G. & G.C. Loury (2005) Affirmative action and Its mythology, Journal of Economic Perspectives, 19-3, 147-162.

[16] Graddy, K. & M. 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.

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

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

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

[20] Kane, T. J. & W.T. William (1998) Racial and ethnic preference in college admissions, in Christopher Jencks and Meredith Phillips (eds.), The Black-White Test Score Gap, Washington: Brookings Institution.

[21] Keith, S., R.M. Bell, A.G. Swanson & A.P. Williams (1985) Effects of affirmative action in medical schools – A study of the class of 1975, The New England Journal of Medicine, 313, 1519-1525.

[22] Knowles, J., N. Persico & P. Todd (2001) Racial bias in motor vehicle searches: theory and evidence, Journal of Political Economy, 109-1, 203-232.

[23] Kobrin, J.L., B.F. Patterson, E.J. Shaw, K.D. Mattern & S.M. Barbuti (2008) Validity of the SAT for predicting first-year college grade point average, College Board, New York.

[24] Kuncel, N. R., S.A. Hezlett & D.S. Ones (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.

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

[26] Ogg , T., A. Zimdars & A. Heath (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.

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

[28] Sackett, P., N. Kuncel, J. Arneson, G. Cooper & S. Waters (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

[29] Sawyer, R. (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

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

Appendix

Part A: Proof of Proposition 1

Consider any feasible rule $p(\cdot)$ satisfying the budget constraint. Since $p^{opt}(\cdot)$ satisfies the budget constraint with equality (recall the definition of γ and q) and $p(\cdot)$ is feasible, we must have

$$\int_{w\in\mathcal{W}} \alpha(w)\, p^{opt}(w)\, dF_W(w) = c \geq \int_{w\in\mathcal{W}} \alpha(w)\, p(w)\, dF_W(w), \tag{10}$$

implying that

$$\int_{w\in\mathcal{W}} \alpha(w) \left[p^{opt}(w) - p(w)\right] dF_W(w) \geq 0. \tag{11}$$

Let $\mathbb{W}(p) := \int_{w\in\mathcal{W}} p(w)\,\alpha(w)\,\beta(w)\, dF_W(w)$. Now, the productivity resulting from $p(\cdot)$ differs from that from $p^{opt}(\cdot)$ by

$$\begin{aligned}
& \mathbb{W}\left(p^{opt}\right) - \mathbb{W}(p) \\
= & \int_{w\in\mathcal{W}} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) + \gamma \int_{w\in\mathcal{W}} \left[p^{opt}(w) - p(w)\right] \alpha(w)\, dF_W(w) \\
\geq & \int_{w\in\mathcal{W}} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) \\
= & \int_{\beta(w)>\gamma} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) \\
& + \int_{\beta(w)<\gamma} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) \\
= & \int_{\beta(w)>\gamma} \left[1 - p(w)\right] \left[\beta(w) - \gamma\right] \alpha(w)\, dF_W(w) + \int_{\beta(w)<\gamma} p(w) \left[\gamma - \beta(w)\right] \alpha(w)\, dF_W(w) \geq 0,
\end{aligned} \tag{12}$$

where the first inequality holds by (11) and that $\gamma > 0$. Therefore, we have $\mathbb{W}\left(p^{opt}\right) \geq \mathbb{W}(p)$ for any feasible $p(\cdot)$, and the solution $p^{opt}(\cdot)$ given in (1) is optimal.

To show the uniqueness, consider any feasible rule $p(\cdot)$ which differs from $p^{opt}(\cdot)$ on some set whose measure is not zero, i.e., $\int_{w\in\mathbf{S}(p)} dF_W(w) > 0$ for $\mathbf{S}(p) := \{w \in \mathcal{W} \mid p^{opt}(w) \neq p(w)\}$. Now, assume that the last equality in (12) holds for this $p(\cdot)$. In this case, since the last equality on the RHS of (12) holds with equality, $p(\cdot)$ must take the following form:

$$p(w) = \begin{cases} 1 & \text{if } \beta(w) > \gamma; \\ 0 & \text{if } \beta(w) < \gamma, \end{cases}$$

for almost every w (with respect to F_W). This implies that $p(w) = p^{opt}(w)$ for almost every w except when $\beta(w) = \gamma$. Since the measure of $\mathbf{S}(p)$ is not zero, we must have $p^{opt}(w) \neq p(w)$ for $\beta(w) = \gamma$, and $\mathbf{S}(p) = \{w \in \mathcal{W} \mid \beta(w) = \gamma\}$, which, together with the budget constraint, implies that $q > p(w)$ when $\beta(w) = \gamma$. However, this in turn implies that we have a strict inequality in the third line on the RHS of (12), which contradicts our assumption. Therefore, we now have shown that $\mathbb{W}\left(p^{opt}\right) > \mathbb{W}(p)$ for any feasible $p(\cdot)$ with $\int_{w\in\mathbf{S}(p)} dF_W(w) > 0$, leading to the desired uniqueness property of $p^{opt}(\cdot)$.

Part B: Evidence of dominance: Other quantiles

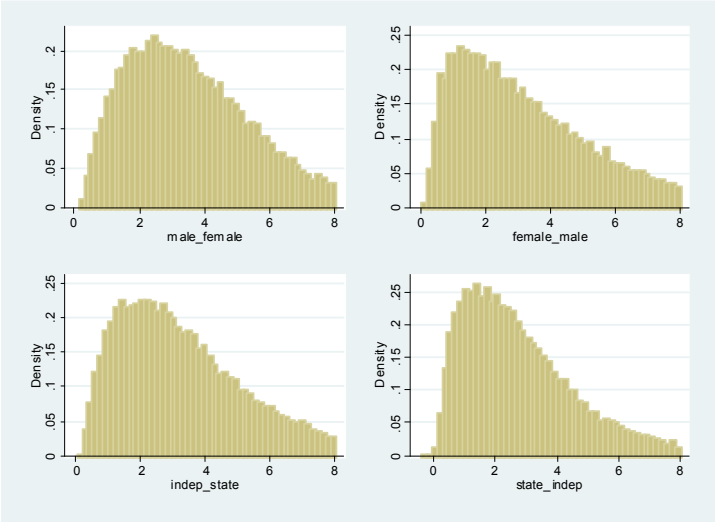


Figure 5: Dominance for 25th percentile

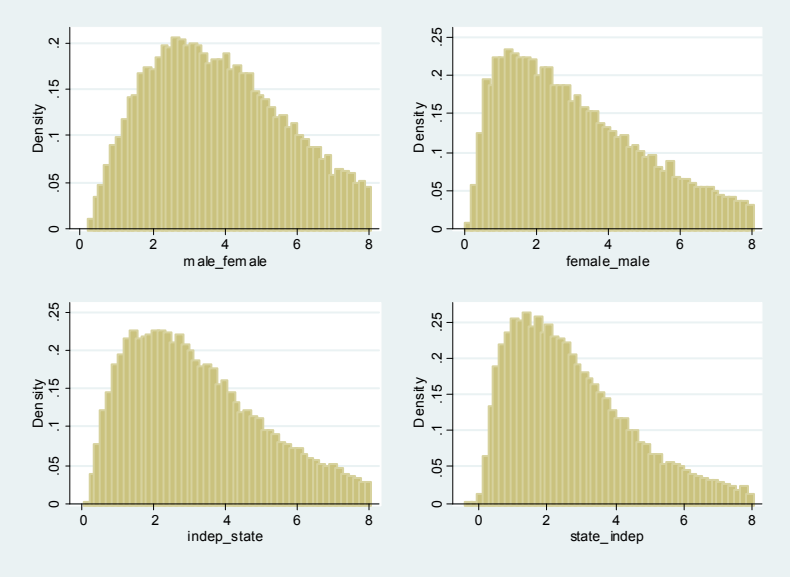


Figure 6: Dominance for 75th percentile

Part C: Test of emptiness

The null hypothesis of an empty $\mathcal{SD}(g,h,\varepsilon)$ can be stated as $\theta_0 \geq 0$, where

$$\theta_0 = \inf_{(x_g,x_h)\in\mathcal{X}_g\times\mathcal{X}_h,\ x_g\succeq_\varepsilon x_h} [p(x_g,g) - p(x_h,h)].$$

The quantity θ_0 is of a form analyzed in Chernozhukov, Lee and Rosen (2013, CLR).[6] We consider constructing a 95% confidence interval for θ_0 in the parametric case $p(x_g,g) = \Phi\left(x_g'\boldsymbol{\delta}_{0h}\right)$ and $p(x_h'\boldsymbol{\delta}_{0h})$ by following the CLR method. Accordingly, denote the dimension of $(\boldsymbol{\delta}_g', \boldsymbol{\delta}_h')'$ by k, a k-variate standard normal by $\mathcal{N}_k$ and the asymptotic variance of $(\hat{\boldsymbol{\delta}}_g', \hat{\boldsymbol{\delta}}_h')'$ by Ω, that is, $\text{AVar}[(\hat{\boldsymbol{\delta}}_g', \hat{\boldsymbol{\delta}}_h')'] = \Omega$. Now the null hypothesis is equivalent to

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

In order to map the notation of this paper into the CLR notation, let

$$\begin{aligned}
v &= (x_g,x_h);\quad \gamma = (\boldsymbol{\delta}_g,\boldsymbol{\delta}_h);\\
\mathcal{V} &= \{(x_g,x_h)\in\mathcal{X}_g\times\mathcal{X}_h:\ x_g\succeq_\varepsilon x_h\};\\
\hat{\theta}_n(v) &= [x_g'\hat{\boldsymbol{\delta}}_g - x_h'\hat{\boldsymbol{\delta}}_h];\\
s_n(v) &= ||(x_g',-x_h')\hat{\Omega}^{1/2}||;\quad Z_n^{\bigstar}(v) = \frac{(x_g',-x_h')\hat{\Omega}^{1/2}}{||(x_g',-x_h')\hat{\Omega}^{1/2}||}\mathcal{N}_k;\\
k_{n,\mathcal{V}}(p) &= Q^p[\sup_{v\in\mathcal{V}} Z_n^{\bigstar}(v)];\\
\hat{\theta}_{n0}(p) &= \inf_{v\in\mathcal{V}}[\hat{\theta}_n(v) + k_{n,\mathcal{V}}(p)\, s_n(v)].
\end{aligned}$$

Then a $100p\%$ one-sided confidence interval (CI) for θ_0 is given by $\hat{C}_n(p) = \left(-\infty, \hat{\theta}_{n0}(p)\right)$. If $\hat{\theta}_{n0}(p) < 0$, then we conclude that $\mathcal{SD}(g,h,\varepsilon)$ is non-empty. In the application, we use $p = 0.95$ and report the CI, $\hat{C}_n(0.95)$, for each choice of g,h.

[6] We note that θ_0 and $\hat{\theta}_n$ (as well as some other components) depend upon upon the choice of $\varepsilon\,(\geq 0)$, but for notational simplicity, we suppress their dependence on ε (the same remark also applies to part D).