AARHUS
UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

CREATES
Center for Research in Econometric Analysis of Time Series

Are University Admissions Academically Fair?

Debopam Bhattacharya, Shin Kanaya
and Margaret Stevens

CREATES Research Paper 2014-6

Department of Economics and Business
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Are University Admissions Academically Fair?

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

February 7, 2014.

Abstract: High-profile universities often face public criticism for undermining academic merit and promoting social elitism through their admissions-process. In this paper, we develop an empirical test for whether access to selective universities is meritocratic. If so, then the academic potential of marginal candidates -- the admission-threshold -- would be equated across demographic groups. But these thresholds are difficult to identify when admission-decisions are based on more characteristics than observed by the analyst. We assume that applicants who are better-qualified on standard observable indicators should on average, but not necessarily with certainty, appear academically stronger to admission-tutors based on characteristics observable to them but not us. This assumption can be used to reveal information about the sign and magnitude of differences in admission thresholds across demographic groups which are robust to omitted characteristics, thus enabling one to test whether different demographic groups face different academic standards for admission. An application to admissions-data at a highly selective British university shows that males and private school applicants face significantly higher admission-thresholds, although application success-rates are equal across gender and school-type. Our methods are potentially useful for testing outcome-based fairness of other binary treatment decisions, where eventual outcomes are observed for those who were treated.

JEL classification: C13, C14, I20, J15.

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

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

** Address: Department of Economics and Business, University of Aarhus and CREATES, Fuglesangs Alle 4, Aarhus V 8210, Denmark. Email: skanaya@creates.au.dk
Kanaya gratefully acknowledges support from CREATES, Center for Research in Econometric Analysis of Time Series, funded by the Danish National Research Foundation (DNRF78).

[***] Address: Department of Economics, University of Oxford, Manor Road Building, Oxford OX1 3UQ, United Kingdom. Email: margaret.stevens@economics.ox.ac.uk

1 Introduction

Admission practices at selective universities generate considerable public interest and political controversy. For example, in the UK a highly publicized 2011 Sutton Trust report shows that nationally just 3% of schools – mostly expensive and independent (as opposed to state-run) institutions – account for 32% of undergraduate admissions to Oxford and Cambridge, while these universities claim to admit solely on the basis of academic merit. On the other hand, background-based admission quotas such as caste-based reservation in India's public universities and race-based affirmative action in American state-funded colleges have been the subject of intense public controversy, the latter recently re-surfacing in the high-profile "Fisher versus University of Texas" lawsuit. Indeed, the question of fair access to selective universities involves two issues of significant interest to economists and policymakers, viz., (a) intergenerational mobility and (b) discrimination – both positive and negative. Despite significant media attention, rigorous analysis of the available micro-evidence on these issues is scant in the literature. In this context, an important starting point would be to directly 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 framework to model such trade-offs, and use it to infer 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). Viewed in this light, if admissions are purely meritocratic, then the marginal admitted student from a state-school should be expected to perform equally well in post-admission assessments (e.g., college exams) as the marginal admit from a private school. But her expected performance would be worse if, say, 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. Conversely, taste-based discrimination against state-schools will lead to the marginal state-school admit to perform better than the marginal independent school admit. The difference between expected performances of marginal candidates across demographic groups can therefore be interpreted as a measure of deviation from meritocracy. A challenge in implementing this approach directly is that a researcher typically observes a subset of the relevant applicant characteristics used by admissions-tutors and the distributions of the unobserved characteristics may – and usually do – differ across demographic groups. This "omitted characteristics" problem jeopardizes the researcher's attempt at reconstructing the decision-maker's perceptions and makes it hard to assess whether the decision-maker acted in an academically unbiased way. Problems of this type been

recognized by previous researchers, especially in the context of detecting taste-based discrimination in labor market hiring; see, for instance, Heckman (1998), Blank et al. (2004) and the references therein. In the present paper, we devise a test for meritocratic admissions – based on the *differences* in admission-thresholds faced by different demographic groups – which is robust to the omitted characteristics problem.

Specifically, we construct an empirical, threshold-crossing model of admissions involving observed applicant covariates and unobserved heterogeneity, i.e., applicant characteristics observed by admission-tutors but unobserved by the researcher. In our model, academic fairness corresponds to using identical thresholds of expected future performance across applicants from different demographic groups. Our key assumption – for which we will provide suggestive empirical evidence – is that applicants who are significantly better in terms of easily observable indicators of academic potential should statistically (but not necessarily with certainty) be more likely to appear stronger to the admission tutor, based on characteristics observed by her but not by the researcher. The distribution of unobservables, conditional on observables, is otherwise allowed to be arbitrarily different across demographic groups. We show that using this assumption in conjunction with pre and post enrolment data, one can learn about the sign and magnitude of the *differences* between admission thresholds applied to different demographic groups. We then apply these methods to analyze admissions data from a popular undergraduate programme of study at a selective UK University. 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, which is not apparent from the equal success rates, thereby illustrating the usefulness of our approach.

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 model and test the extent of meritocracy – in Becker's sense – of college admissions. The present paper attempts to fill this gap. In particular, it focuses on the *marginal* admits in different demographic groups and thereby shows that equal success rate in admissions across demographic groups can be – and indeed is in our application – consistent with very different admission standards across these different groups. This 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). Our paper also complements an existing literature on analyzing the *consequences* of affirmative actions in college admissions. Fryer and Loury (2005) provide a critical review of the relevant theoretical literature and a comprehensive bibliography. On the empirical side, Arcidiacono (2005) uses a structural model of admissions to simulate the potential, counterfactual consequences of removing affirmative action in US college admission 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 the present paper, we construct a formal econometric model where affirmative action and meritocracy have contradictory empirical implications, and uses it in conjunction with admissions-related micro-data to detect deviations from meritocracy in prevalent admission practises.

The rest of the paper is organized as follows: Section 2 sets up a simple theoretical model; Section 3 lays out 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; 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 (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 expected 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:*

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

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

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

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

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

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

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*). Proof in appendix.*

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.[2]

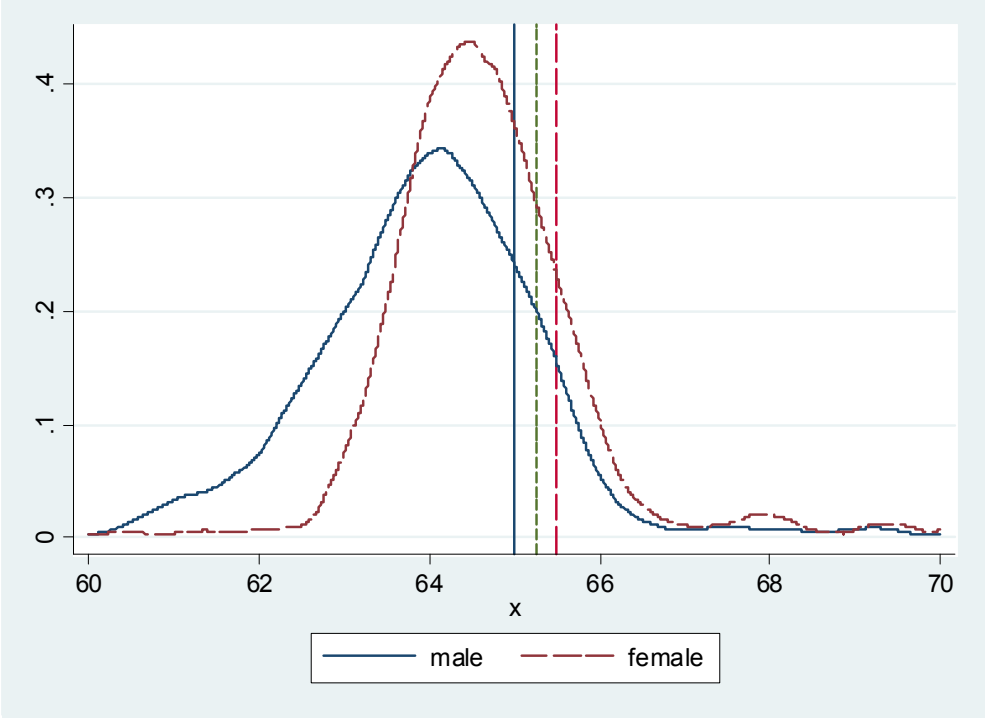

Figure 1: Equal threshold versus equal probability of acceptance

In this graph, the solid curve represents the marginal density of academic merit for male applicants and the dashed curve that for female applicants. Under identical thresholds, marked by the small-dashed 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 dashed density curve for female applicants. The graph shows that the latter area is significantly larger, suggesting that if a common threshold were used, admission rate for female applicants would be higher. Conversely, equating admission probabilities across gender requires employing a larger threshold (marked by long dash) for females than for males (solid line). 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 admission-tutors' inference about academic merit were based on more characteristics than we the researchers observe, so that we cannot replicate the two density curves as in the previous graph. Therefore, we now turn to the task of constructing an econometric model incorporating unobserved heterogeneity in an empirical model of admissions.

[2]The first author is grateful to Amitabh Chandra and Doug Staiger for suggesting this illustration.

3 Econometric Model

To set up the 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 and one or more past years. In addition, we have data on outcomes (e.g college GPA) for those past applicants who had enrolled. 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. 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 support of X^P conditional on $G^P = g$ and $A^P = 1$, i.e.,

$$\mathcal{X}_g^P := \left\{x : \Pr\left[A^P = 1 | X^P = x, G^P = g\right] > 0\right\}.$$

This is the set of the values of X^P which occur among the admits of type g in past years and so one can, in principle, calculate (i.e., 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 (such as reference letters) which are *unobservable* to the analyst but observed by the admission-tutor, e.g., reference letters. 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.

Remark 1 *Note that the interpretation of Z is not that it is the level of unobserved characteristics themselves; rather, it is the applicant-quality inferred from such attributes. For example, if teachers at private schools are better trained to write reference-letters, then admission-tutors are expected to take this into account when forming their impression Z.*

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 i's academic merit, e.g., how applicant i will perform when admitted. In accordance with our economic model, we assume that an applicant i with $G_i = g$, $Z_i = z$ and $X_i = x \in \mathcal{X}_g$ is offered admission (i.e., $D_i = 1$) if and only if $\mu_i^* = \phi(x, g, z) \geq \gamma_g$, where μ_i^* denotes the subjective conditional expectation of applicant i's academic potential 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,

$$D_i = \begin{cases} 1 & \text{if } \phi(X_i, G_i, Z_i) \geq \gamma_{G_i}; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

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 perceived academic merit. Having a larger γ for, say, females than males implies that a male applicant with the same expected outcome as a female applicant is more likely to be admitted. Conversely, under affirmative action type policies, γ_g will be lower for those gs which represent historically disadvantaged groups. Therefore, we are interested in testing whether the values of the threshold γ_g 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 female applicant with identical X as a male candidate can have a higher probability of being admitted and yet the admission process may be academically fair if females have a higher expected outcome than males with identical X.

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 GPA and SAT for the entire population of applicants. We will denote by $Q^\tau(Z|A)$ the τth quantile of the random variable Z given the random variable A.

[2] 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).

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) $\mu_i^* = \phi\left(X_i, G_i, Z_i\right)$ (introduced just before equation (2)) is continuously distributed conditionally on any realization of (X_i, G_i).

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

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

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

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

Discussion: Crudely speaking, Assumption M/SD means that applicants who are better along standard, observable indicators of academic ability are also likely to be better – "on average" – in terms of the index of unobserved characteristics which the tutors weigh positively in determining admissions. The motivation for this assumption comes from the fact that for meritocratic admissions, the outcome of interest may be thought of as a measure of future academic performance whereas the measures in X are a set of past academic performance in high-school or admissions-related assessments. It is therefore likely that candidates who have performed significantly better in past assessments are statistically more likely to have performed better in those assessments (unobserved by the researcher) which admission tutors view as positive determinants of future performance and hence, under the assumption of being academically motivated, would weigh positively in the decision to admit.

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 $x_g \succeq_\varepsilon x_h$ for a larger ε 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. When we use an $\varepsilon > 0$, rather than $\varepsilon = 0$, our identifying information for admission-thresholds will come from pairs of applicants who are "well-separated" in terms of their prior test-scores. In the application, we will use values of $\varepsilon = 0.1$ and $\varepsilon = 0.25$ which are strictly positive and thus lead to comparison of applicant-pairs with no overlap of pre-admission test-scores. Pairs who are very close to each other in terms of observables are not used in the analysis.

Assumption M is substantively much weaker than two informal arguments often used in applied work – viz., (i) when the distribution of the observable covariates are balanced across treatment and control groups in quasi-experimental designs, it is taken to imply that they are also balanced in terms of unobservables (e.g., Greenstone and Gayer, 2009) and (ii) orthogonality of an instrument with observed covariates is taken as suggestive evidence that it is orthogonal with unobserved covariates (e.g., Angrist and Evans, 1998, p. 458). In our context, the type of variables typically unobservable to researchers but likely to affect admissions include achievements such as winning special academic prizes, participation in science or math olympiads, high intellectual enthusiasm conveyed by applicants' personal essays and the subjective impressions of previous teachers implied via reference letters. Such specific information can identify individual applicants and therefore are most likely to be withheld from researchers owing to privacy considerations. However, while making 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 also remark 1). See Section 6 below for evidence that is suggestive and supportive of this assumption, in our application.

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

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

Remark 3 *The restriction on the median cannot be replaced by a restriction on the conditional expectation for identification purpose since we are considering a discrete-choice problem, viz., $D = \mathbf{1}\{\phi(X, G, Z) \geq \gamma_G\}$. See Manski (1975) for why a conditional quantile restriction is necessary for the identification of discrete-choice models.*

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

4.1 Sign of threshold differences

We first show how assumption M or SD can help identify the sign of threshold differences across demographic groups. To do this, we impose an intuitive assumption on the structure of the ϕ function.

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

Discussion: Part (i) of Assumption CM is essentially definitional (regarding Z) in that higher values of the index of ability based on unobserved characteristics are associated with higher values of the perceived expected outcome. Part (ii) says that if a g-type applicant is better than an h-type applicant along a set of key observable characteristics *and* is at least equally good along the ability index which is unobservable to us but observable to the decision-makers, then the g-type applicant will be perceived to have a higher expected outcome by the decision-maker. It is important for part (ii) that the g-type applicant is at least as good as the h-type applicant along the index Z; without this condition, it is easy to come up with counter examples. For instance, suppose that admission tutors base their assessment on past written exams whose scores X are observed by us (researchers) and the quality of the reference letter Z, unobserved by us. Then a female candidate who has scored lower on every component of X than a male candidate but has a much better recommendation may or may not be perceived as having a lower potential than the male candidate. But a female candidate who has an equally strong recommendation Z as a male candidate but has scored lower on every X than him will likely be perceived to have lower academic potential (note also remark 1) in expectation.

Now, assumptions M and CM can be used to identify the sign of threshold differences. To see this, define the function

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

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

$$\mathcal{M}(g,h,\varepsilon) := \{(x_g, x_h) \in \mathcal{X}_g \times \mathcal{X}_h : x_g \succeq_\varepsilon x_h,\ p(x_g, g) \leq 0.5 < p(x_h, h)\}. \tag{3}$$

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

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

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

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

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

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

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

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

since

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

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

Note that so far we have not used any information on post-admission performance of applicants and not taken any stance on what the ultimate measure of academic merit is. We have only assumed that the observed covariates X are used by admission tutors to infer an overall measure of academic potential. Thus the signs of threshold differences obtained above are valid under any expectation formation process (i.e., $\phi(\cdot, \cdot, \cdot)$), as long as assumptions CM are satisfied.

4.2 Magnitude of threshold difference

In order to infer the extent of deviation from meritocracy (over and above its direction), we need to specify a post-enrolment outcome as the relevant measure of academic merit. Accordingly, we now assume that post-entrance exam performance (e.g. final GPA) is the relevant outcome.[3] Accordingly, define

$$\mu^P(x, g) = E\left[Y^P | X^P = x, G^P = g, A^P = 1\right], \tag{5}$$

the conditional expectation of outcome Y^P for a past enrolled applicant given his/her characteristics $(X^P, G^P) = (x, g)$ and impose the following stronger (than CM) assumption on the structure of $\phi(X_i, Z_i, G_i)$.

Assumption AS (Additive Separability) The tutors' subjective assessment ϕ satisfies

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

[3] Indeed, one may use any other post-enrolment outcome which is observed for all enrolled students, e.g., finishing the program, salary upon graduation etc. and define meritocracy in terms of that outcome.

where $\mu^P(x,g)$ is defined in (5).[4, 5]

Discussion: Assumption AS also 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 $\mu^P(X_i,G_i)+Z_i$.[6]

Assumptions AS and M yield a lower bound on the *magnitude* of threshold differences. To see this, note that

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

This implies that

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

[4] Note from (5) 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. 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.

[5] 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.

[6] Strictly speaking, Assumptions AS and CM are non-nested in that the former does not require the "monotonicity" in x for fixed z while Assumption CM does not require the additively separable structure. On the other hand, monotonicity is quite natural in this context and thus CM is a substantively weaker assumption.

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 $(x_g, x_h) \in \mathcal{M}(g, h, \varepsilon)$ 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

$$\gamma_g - \gamma_h \geq \sup_{(x_g,x_h)\in\mathcal{M}(g,h,\varepsilon)} \left[\mu^P(x_g, g) - \mu^P(x_h, h)\right] \equiv \underline{\theta}(g, h). \tag{6}$$

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 $p(\cdot, \cdot)$. Thus we obtain a lower bound on the magnitude of threshold differences in addition to its sign. Under the stronger condition of Assumption SD, we have the bound

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

where $\mathcal{SD}(g, h, \varepsilon)$ is defined in (4).

Intuitively speaking, here the identification-relevant information also 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 also have higher predicted performance based on observables, then they must have been facing a higher threshold leading to a lower probability of admission.

Some Alternative Identification Strategies: In the healthcare context, 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 the 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. In contrast, our analysis imposes no such restriction on the unobservables' distribution. In a healthcare application, 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. For law-enforcement and healthcare provision, several researchers have used reasoning based on economic optimization by the subjects to detect racial prejudice (c.f. Persico, 2009 for a survey). However, these approaches rely on the specifics of the context and do not generalize to situations involving university admissions. For example, it is both difficult for university-applicants to alter their potential academic outcomes in response to admission protocols and impractical for them to want to do this, given the one-shot nature of admission exercise.

5 Estimation and Inference

Given the identification analysis above, our next task is to develop a formal inference method for testing threshold-differences. For this purpose, we will make the stronger assumption of SD, rather than M. Indeed, these two assumptions have the same intuitive interpretation; the evidence for SD (see section 6 and and also part B of the Appendix) is strong and conducting statistical inference under it is slightly simpler.

The key task regarding inference – corresponding to Assumptions SD and CM – is to test whether $\mathcal{SD}(g,h,\varepsilon)$ defined in equation (4), viz.,

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

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

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

We will now outline how to test the emptiness of $\mathcal{SD}(g,h,\varepsilon)$, based on an inference method developed for "intersection bounds" by CLR (2013). Although our identification method is non-parametric in the sense of not requiring functional form specifications, estimation and inference for

the nonparametric case is complicated. Due to relatively small sample-size, the two-sample nature of the problem and the complicated construction of "intersection bounds" for nonparametric estimates (requiring subjective choice of various tuning parameters), we do not consider such methods here. Instead, we focus on the case where $p(\cdot,\cdot)$ is parametrically specified as a probit.[7] That is,

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

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

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

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

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

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

Inference on the magnitude bounds: When bounding the *magnitude* of threshold differences as described in subsection (4.2), we consider a slightly more conservative bound which is easier to conduct inference on. Note from (6) that the key parameter of interest is a supremum over the domain $\mathcal{SD}(g, h, \varepsilon)$, defined in (4). 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 (6)) by the upper λth quantile, and conduct inference on it. That is, we use the *implication* of (6) 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 (6):

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

[7] We take the supports $\mathcal{X}_g$ and $\mathcal{X}_h$, to be known.

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 a *two-sample* moment condition problem where the moments are nonsmooth in the parameters.[8] 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, one 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 a competitive and popular undergraduate degree programme 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

[8] Denote the two sets of observations associated with g and h, by $\{X_{g,i}\}_{i=1}^{n_g}$ and $\{X_{h,j}\}_{j=1}^{n_h}$,respectively; let $n = n_g + n_h$; and let n be the total number of observations in the current cohort. Then our estimator $\hat{\theta}^\lambda = \hat{\theta}^\lambda(g,h)$ is the solution in θ to the sample moment equation:

$$M_n^\lambda(\theta, \hat{\boldsymbol{\alpha}}) = 0,$$

where $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\beta}}_g', \hat{\boldsymbol{\beta}}_h', \hat{\boldsymbol{\delta}}_g', \hat{\boldsymbol{\delta}}_h')'$; $(\hat{\boldsymbol{\beta}}_g, \hat{\boldsymbol{\beta}}_h)$ and $(\hat{\boldsymbol{\delta}}_g, \hat{\boldsymbol{\delta}}_h)$ are first-step estimates of linear regression and probit models, respectively (the former is based on the past cohorts' data and the latter is on the current cohort'); and

$$M_n^\lambda(\theta,\alpha) := \frac{1}{n_g n_h}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left[\lambda - \mathbf{1}\{X_{g,i}'\boldsymbol{\beta}_g - X_{h,j}'\boldsymbol{\beta}_h \leq \theta\}\right]\mathbf{1}\{X_{g,i} \succeq_\varepsilon X_{h,j}, X_{g,i}'\boldsymbol{\delta}_g \leq X_{h,j}'\boldsymbol{\delta}_h\}. \tag{10}$$

From (10), it transpires that our estimator is a two-sample moment based estimator, where the sample moment conditions are nonsmooth in the parameters but the population versions (upon taking expectations) are.

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-domiciled applicants. The application process consists of an initial stage whereby a standardized "UCAS" form is filled by the applicant and submitted to the university. This form contains the applicant's unique identifier number, gender, school type, prior academic performance record, personal statement and a letter of reference from the school. The 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 the aptitude test 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 candidate-specific information from among the applicants 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.

Choice of covariates: We chose a preliminary set of potential covariates to be the observables, based on the information recorded on UCAS forms and the university's application records. We use as observable components (i.e.,X) the 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 (distinct from 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. They perform slightly worse on average in their GCSE and A-levels. These differences are statistically significant at the 5% level. Note that there is no significant difference in offer rates between male and female candidates. The independent and state school applicants are quite similar in terms of most characteristics except for a slightly higher GCSE for the former.

In Table 2 we report the results of a probit regression of receiving an offer across all applicants. Table 2 strengthens the findings from Table 1 by showing that even after controlling for covariates, gender and school-type do not affect the *average* admission-success rate among applicants. The value of McFadden's pseudo-R^2 for the probit model is about 50% and the corresponding R^2 for a linear probability model (not reported here) is about 45% – which are about 10 times higher than the goodness-of-fit measures typically reported by applied researchers working with cross-sectional data. This suggests that the commonly observed covariates explain a very large fraction of admission outcomes. Moreover, Table 2 also shows that the aptitude test and interview scores have the largest impact upon receiving an offer for the applicant population (in terms of the t-statistics).

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 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.

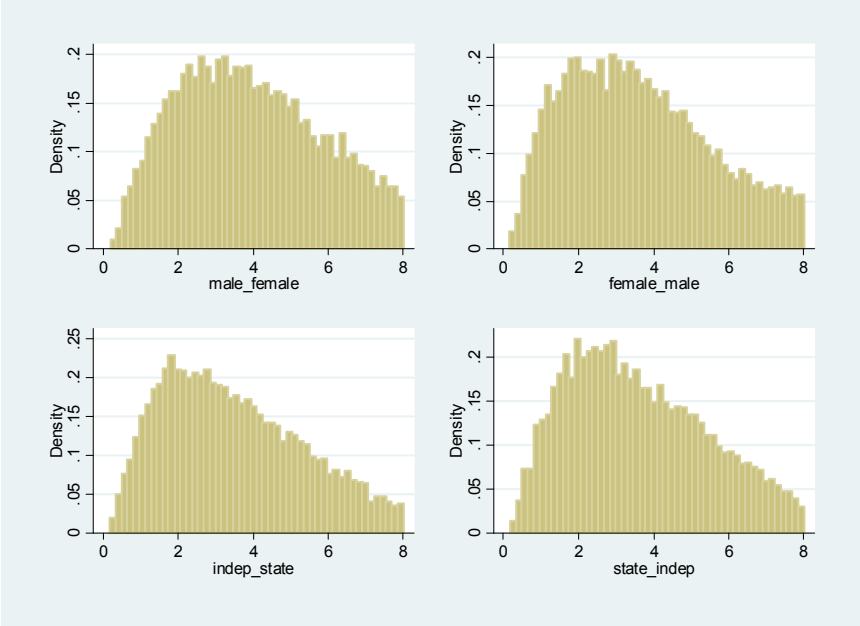

Figure 2: Evidence of Median Dominance

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

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. The purpose is to investigate how well our method works when we a priori know the admission thresholds. In order implement 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 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 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 (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

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

where $0.05Z_i$ 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 $0.05Z_i + u_i$ 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 (see eqn (8)) with $\lambda = 0.5$ (median), $\lambda = 0.80$ as well as the mean, using $\varepsilon = 0.1$. The table can be read as follows. The first column reports the true value of δ, the second column shows the upper limit of the one-sided confidence interval for testing emptiness of $\mathcal{SD}(g, h, \varepsilon)$. 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 a nonempty $\mathcal{SD}(g, h, \varepsilon)$ and positive point-estimates for δ but the associated confidence intervals contain 0. For a negative value of δ, an empty $\mathcal{SD}(g, h, \varepsilon)$ cannot be rejected, as expected. Overall, this table presents strong evidence that our method works well in practice.

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 is assumed to dominate that for the latter for all g and h, as per Assumption M or SD above.

In accordance with the discussion in Section 5 the first step is to examine emptiness of $\mathcal{SD}(g, h, \varepsilon)$ using data on only X and D. We first investigate this graphically by plotting the marginal C.D.F. of the difference in admission probabilities $p(X_g, g) - p(X_h, h)$ for pairs of (X_g, X_h) satisfying $X_g \succeq_\varepsilon X_h$ for $\varepsilon = 0.1$ for various combinations of g and h.[9] When the event $\{X_g \succeq_\varepsilon X_h\}$ happens with positive probability, an empty $\mathcal{SD}(g, h, \varepsilon)$ is equivalent to $\Pr[X_g \succeq_\varepsilon X_h,\ p(X_g, g) < p(X_h, h)] = 0$,

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

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.

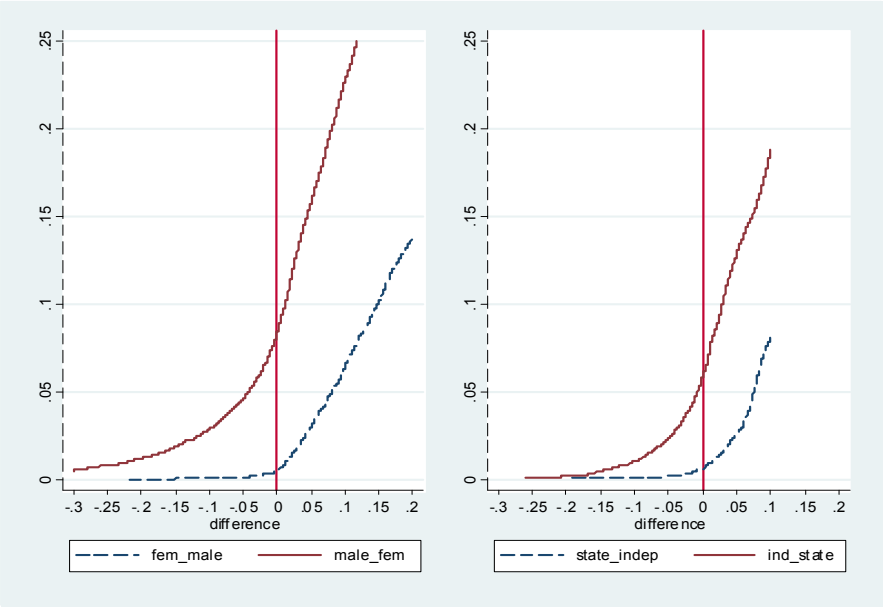

Figure 3: Evidence of Emptiness

Clearly, the first curve has significant mass below zero and the dashed curve has almost no mass below zero, suggesting a positive probability that $p(X_{male}, male) < p(X_{female}, female)$ although $X_{male} \succeq_{\varepsilon} X_{female}$. This evidence is still present in the right panel with independent and state schools replacing male and female, respectively, but to a slightly lesser extent, suggesting that γ_{indep} may be only slightly larger than γ_{state}. To perform the test formally, in Table 4 column 2, we report $\hat{\theta}_{0n}(0.95)$, the upper limit of a one-sided confidence interval, calculated using the method of CLR, as explained in Section 5. A negative upper limit indicates that the set $\mathcal{SD}(g, h, \varepsilon)$ is nonempty and consequently we reject the null of $\gamma_g \leq \gamma_h$ in favour of $\gamma_g > \gamma_h$. It is evident from 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. This clearly suggests that males and private school applicants face higher admission thresholds.

The exact upper limits of confidence intervals reported in Table 4 vary slightly across functional specifications (e.g. whether higher order terms and interactions in the test scores are or are not used to estimate $p(\cdot, \cdot)$), but two empirical findings are robust across all specifications: (a) the gender gap is large, persistent and statistically significant in every case, and (b) the independent-state

school difference is comparatively smaller.

Given the conclusion of the test of emptiness, we now impose Assumption AS and 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 $\varepsilon = 0.25$. In Tables 5A and 5B we report the estimated lower bounds $\hat{\theta}^{\lambda}$ for $\lambda = 0.80$, given by (6) and (7), using prelim and finals performance as outcomes, respectively. The first column, labeled "upper limit", reports $\hat{\theta}_{0n}(0.95)$ from the previous table. When this number is negative, it indicates that the $\mathcal{SD}(g, h, \varepsilon)$ is nonempty, whence we proceed to compute $\hat{\theta}^{\lambda}$. Under Assumptions SD and CM, a nonempty $\mathcal{SD}(g, h, \varepsilon)$ already indicates that $\gamma_g > \gamma_h$. Upon imposing the substantively stronger assumption of 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.

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.

Interpretation of the empirical findings: It would be natural to conjecture that the threshold differences arise primarily from the implicit or explicit practice of affirmative action, viz., the overweighting of outcomes for historically disadvantaged groups. A second possibility is that, in

face of political and/or media pressure, admission tutors try to equate an application success rate for, say, males with one for females, which is also consistent with our empirical findings (see Tables 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 and independent school applicants face significantly higher admission thresholds.

In order to gain some further insight into how the threshold discrepancies arise, we plot the empirical C.D.F.s of predicted academic performance based on the observable characteristics. This is done by regressing first-year examination scores in university on aptitude test and essay score, interview grades and gender/schooltype for enrolled students. The estimated CDFs of predicted performance by gender (the left panel) and by schooltype (the right panel) are plotted in fig. 4. It is clear that the male distribution first-order stochastic dominates the female distribution. This means that if admissions were determined solely by predicted performance based on observables (i.e., there is no unobserved heterogeneity), *any* common acceptance rate across gender will result in a higher predicted outcome for the marginal accepted male than the marginal accepted female.

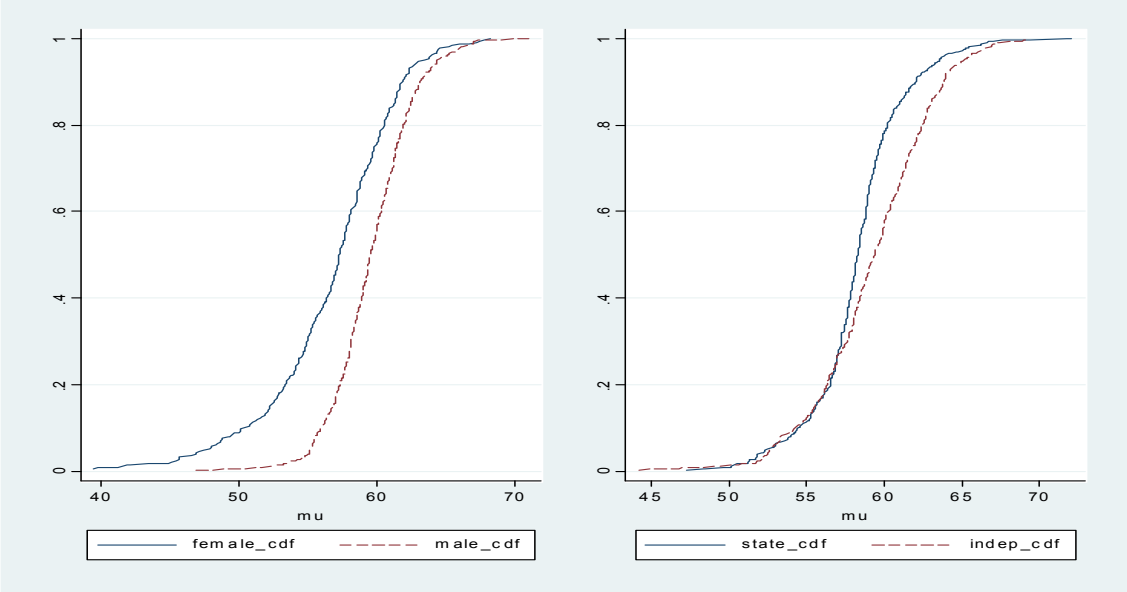

Figure 4: Source of Differences

This can be seen in Figure 4, by looking along any fixed cutoff on the vertical axis. Any such

horizontal cut-off line[10] 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. A similar, albeit relatively weaker, dominance situation occurs for school-type, as can be seen in the right-hand graph in Figure 4. Our results in table 4 imply that allowing for unobserved heterogeneity does not change this scenario substantively, and suggests that equating the application success-rates (see table 1) leads to the use of higher admission thresholds for male and for private school candidates.

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

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

[10] 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.

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

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

where $\mathcal{SD}(g,h,\varepsilon)$ is defined in (4). 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 (6) 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 a rigorous empirical approach to testing, on the basis of micro-data, whether and to what extent an existing admission protocol is efficient, i.e., meritocratic, when a researcher observes some but not all applicant-specific information observed by admission tutors. The approach works by obtaining the sign and lower bounds on the magnitude of *difference* in admission thresholds faced by applicants of different demographic groups. These quantities are robust to the unobserved characteristics problem, under an intuitive assumption about the ranking of applicants by unobservable attributes. They 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. 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. Our methods are potentially useful for testing outcome-based fairness of binary decisions 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
aptitude test	Overall score in Aptitude Test 0-100
essay	Score on Substantive Essay 0-100
Interview	Performance score in interview 0-10(
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.

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
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
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	RHS of CLR CI, indep-state	Median	Mean	80%ile
4	-4.14	3.32	3.35	4.02
		(1.35, 4.93)	(1.59, 5.02)	(2.21, 6.21)
3	-4.45	1.92	1.88	2.63
		(0.28, 3.22)	(0.05, 3.01)	(1.96, 4.07)
2	-3.57	1.33	1.31	1.54
		(0.76, 2.78)	(0.28, 2.31)	(0.51, 2.88)
1	-1.66	0.86	0.86	1.21
		(0.14, 1.60)	(-0.08, 1.36)	(-0.88, 1.99)
0	0.13	-0.06	0.11	0.29
		(-1.88, 0.47)	(-1.78, 0.47)	(-1.45, 0.61)
-2	1.86	.	.	.

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 right limit of the one-sided confidence interval for testing emptiness with a negative value indicating emptiness.. 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	upper limit of CI for testing 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	upper limit of CI for testing 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 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] Bhattacharya, D. & P. Dupas (2012) Inferring efficient treatment assignment under budget constraints, Journal of Econometrics, 167, 168-196.

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

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

[8] 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.

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

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

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

[12] 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.

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

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

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

[24] 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

[25] 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

[26] Tamer, Elie (2010): Partial Identification in Econometrics," Annual Reviews of Economics, Vol. 2, No.1, 2010, pp. 167-195.

[27] 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.

Technical 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

$$\int_{w\in\mathcal{W}} \alpha(w)\, p^{opt}(w)\, dF_W(w) = c \geq \int_{w\in\mathcal{W}} \alpha(w)\, p(w)\, dF_W(w), \tag{11}$$

implying that

$$\int_{w\in\mathcal{W}} \alpha(w) \left[p^{opt}(w) - p(w)\right] dF_W(w) \geq 0. \tag{12}$$

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

$$\begin{aligned}
& \mathbb{W}\left(p^{opt}\right) - \mathbb{W}(p) \\
= \; & \int_{w\in\mathcal{W}} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) + \gamma \int_{w\in\mathcal{W}} \left[p^{opt}(w) - p(w)\right] \alpha(w)\, dF_W(w) \\
\geq \; & \int_{w\in\mathcal{W}} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) \\
= \; & \int_{\beta(w)>\gamma} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) \\
& + \int_{\beta(w)<\gamma} \left[p^{opt}(w) - p(w)\right] \alpha(w) \left[\beta(w) - \gamma\right] dF_W(w) \\
= \; & \int_{\beta(w)>\gamma} \left[1 - p(w)\right] \left[\beta(w) - \gamma\right] \alpha(w)\, dF_W(w) + \int_{\beta(w)<\gamma} p(w) \left[\gamma - \beta(w)\right] \alpha(w)\, dF_W(w) \geq 0,
\end{aligned} \tag{13}$$

where the first inequality holds by (12) 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 (13) holds for this $p(\cdot)$. In this case, since the last equality on the RHS of (13) 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 (13), 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

The following histograms are for substantiating assumption $\mathcal{SD}$. They are analogous to those reported in figure 2 but for quantiles other than the median. For example, the top left histogram in Fig. 5 corresponds to

$$\mathrm{Q}^{.25}[interview \mid X_{male}, male] - \mathrm{Q}^{0.25}[interview \mid X_{female}, female]$$

computed across all pairs of males and females satisfying $X_{male} \succeq_{\varepsilon=0} X_{female}$. The strictly positive support of these histograms implies dominance with respect to quantiles other than the median.

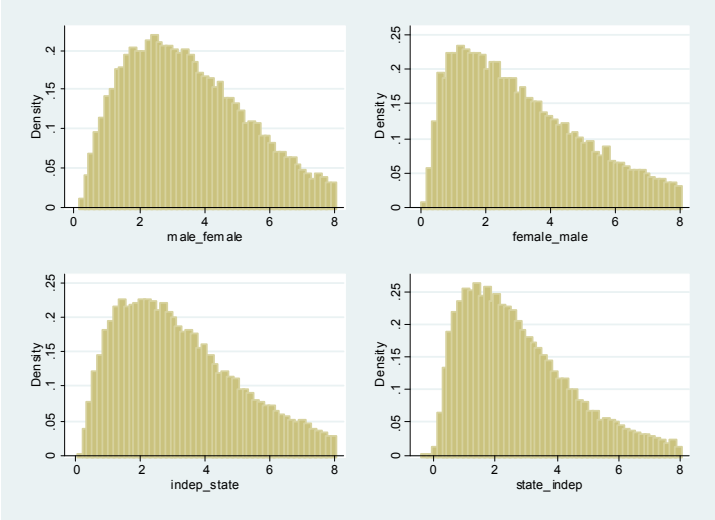

Figure 5: Dominance for 25th percentile

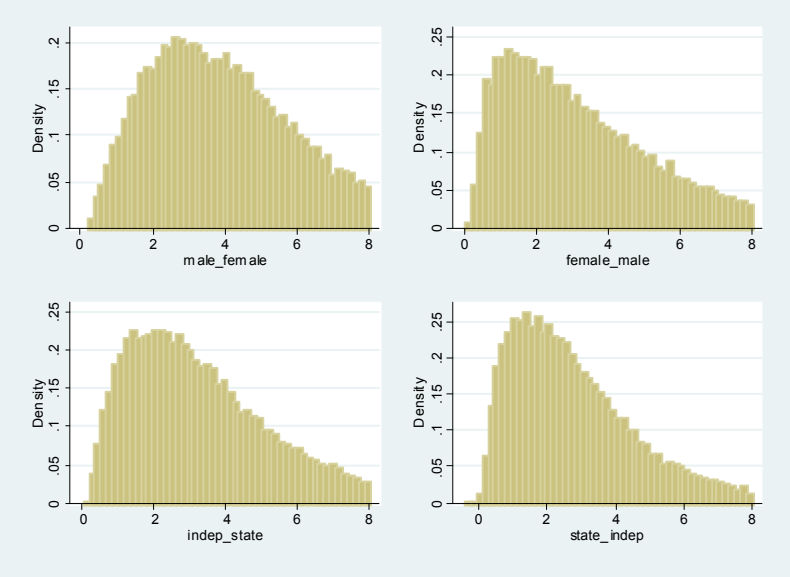

Figure 6: Dominance for 75th percentile

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). 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, $\text{AVar}[(\hat{\boldsymbol{\delta}}_g', \hat{\boldsymbol{\delta}}_h')'] = \Omega$. Denote the τth quantile of a random variable W by $Q^\tau(W)$. 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.

Part D: Asymptotic distribution of $\hat{\theta}^{\lambda}(g,h)$ [Technical Appendix for Online Publication]

First, we outline some main points for our derivation of the asymptotic distribution of the quantile-based lower bound estimator $\hat{\theta}^{\lambda}(=\hat{\theta}^{\lambda}(g,h))$ of $\theta_0^{\lambda}(=\theta_0^{\lambda}(g,h))$.[1] Then, we formally present a set of conditions required for our asymptotic analysis, our obtained result, and its proof.

Outline of asymptotic analysis: As stated earlier in Section 5, $\hat{\theta}^{\lambda}(g,h)$ is defined as a solution to the sample moment condition $M_n^{\lambda}(\theta,\hat{\boldsymbol{\alpha}})=0$, where $\hat{\boldsymbol{\alpha}}$ is a vector containing preliminary probit and linear-regression estimators. In handling this moment condition, we are faced with two non-standard problems: The first one is the non-smoothness of $M_n^{\lambda}(\theta,\boldsymbol{\alpha})$. Since this objective, defined in (D.8), is computed based on indicator functions, it is neither continuous nor differentiable, which makes it impossible to use standard Taylor-expansion arguments. However, we can check the (continuous) differentiability of the limit $\bar{M}^{\lambda}(\theta,\boldsymbol{\alpha})$ of $M_n^{\lambda}(\theta,\boldsymbol{\alpha})$, and apply the Taylor expansion to $\bar{M}^{\lambda}(\hat{\theta}^{\lambda},\hat{\boldsymbol{\alpha}})$ (instead of $M_n^{\lambda}(\hat{\theta}^{\lambda},\hat{\boldsymbol{\alpha}})$), which is a usual trick as in Pakes and Pollard (1989, PP) and Chen, Linton and Keilegom (2003, CLK) and which guarantees the $\sqrt{n}$ asymptotic normality. The second problem is that M_n^{λ} is computed based on two samples $\{X_{g,i}\}$ and $\{X_{h,j}\}$. For results as in PP or CLK, a key is the so-called stochastic equicontinuity property (SEP) of an objective function. Various results and techniques for verifying this property are available in the standard one-sample case, as found in Andrews (1994). However, such results have not been well-established in the literature for two-sample cases like ours, and we below derive the SEP of a normalized objective: $\nu_n(\theta,\boldsymbol{\alpha})=\sqrt{n}[M_n^{\lambda}(\theta,\boldsymbol{\alpha})-\bar{M}^{\lambda}(\theta,\boldsymbol{\alpha})]$.

To obtain further insights on the development of asymptotic theory, look at the following key decomposition:

$$\begin{aligned}
0 &= \sqrt{n}M_n^{\lambda}(\hat{\theta}^{\lambda},\hat{\boldsymbol{\alpha}})=\sqrt{n}[M_n^{\lambda}(\hat{\theta}^{\lambda},\hat{\boldsymbol{\alpha}})-\bar{M}^{\lambda}(\theta_0^{\lambda},\boldsymbol{\alpha}_0)] \\
&= \sqrt{n}[\bar{M}^{\lambda}(\hat{\theta}^{\lambda},\hat{\boldsymbol{\alpha}})-\bar{M}^{\lambda}\left(\theta_0^{\lambda},\boldsymbol{\alpha}_0\right)]-\sqrt{n}[M_n^{\lambda}(\theta_0^{\lambda},\boldsymbol{\alpha}_0)-\bar{M}^{\lambda}(\theta_0^{\lambda},\boldsymbol{\alpha}_0)] \\
&\quad +\sqrt{n}[M_n^{\lambda}(\hat{\theta}^{\lambda},\hat{\boldsymbol{\alpha}})-\bar{M}^{\lambda}(\hat{\theta}^{\lambda},\hat{\boldsymbol{\alpha}})+M_n^{\lambda}(\theta_0^{\lambda},\boldsymbol{\alpha}_0)-\bar{M}^{\lambda}(\theta_0^{\lambda},\boldsymbol{\alpha}_0)], \qquad \text{(D.2)}
\end{aligned}$$

where we note that $\bar{M}^{\lambda}(\theta,\boldsymbol{\alpha})=E[M_n^{\lambda}(\theta,\boldsymbol{\alpha})]$ and $\bar{M}^{\lambda}\left(\theta_0^{\lambda},\boldsymbol{\alpha}_0\right)=0$; the first term on the RHS can be further expanded by the Taylor expansion, which may be called as a delta-method term; and the second term may be called as a central limit theorem (CLT) term while the third may be called as a stochastic-equicontinuity one. As inferred from its name, the second term gives the limit normal distribution. Since M_n^{λ} involves double summations, a standard

[1] We note that $\theta_0^{\lambda}(g,h)$, $\hat{\theta}^{\lambda}(g,h)$, $\bar{M}^{\lambda}(\theta,\boldsymbol{\alpha})$ and $M_n^{\lambda}(\theta,\boldsymbol{\alpha})$ (and some other components) depend upon the choice of $\varepsilon\,(\geq 0)$, but for notational simplicity, we suppress their dependence on ε.

CLT cannot be directly applied. However, M_n^λ admits a projection-based decomposition (as U and V statistics) and can be asymptotically expressed as the sum of two (normalized) summations each of which is based on one sample, to which the standard CLT can be applied. The third term on the RHS of (D.2) is shown to be asymptotically negligible. This is proved by using the SEP of $\nu_n(\theta, \boldsymbol{\alpha})$ (Lemma 3) as well as the convergence of $(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}})$ to $\left(\theta_0^\lambda, \boldsymbol{\alpha}_0\right)$ (Lemmas 1 and 2).

It is not trivial to verify the SEP of $\nu_n(\theta, \boldsymbol{\alpha})$. Note again that the object M_n^λ involves double summations. In this respect, it has some similarity to U and V statistics. However, results for such statistics (e.g., Nolan and Pollard, 1987, 1988; Sherman, 1993, 1994; Section 8.2 of Newey and McFadden, 1994) are not directly applicable to our case, since our M_n^λ is based on two samples but U and V statistics are computed using pairs (or higher tuples) drawn from a single sample. Therefore, we present the SEP of $\nu_n(\theta, \boldsymbol{\alpha})$ and a related uniform convergence (UC) result in Lemma 3, and develop an independent proof, which is based on a covering-number technique (from empirical process theory) and the Bernstein exponential inequality. The result of the lemma may be the most comparable to Sherman's U-statistic results (in particular Theorem 3 of Sherman, 1993 and a set of results in Sherman, 1994). Although we exploit the fact that a set of functions concerned is Euclidean as in Sherman (1993, 1994) (as well as in Nolan and Pollard, 1987, 1988), our proof is different form theirs. While they use a sort of symmetrization technique, we use the Bernstein inequality, by which we can exploit the independence between two samples – direct use of such inequality may not be necessarily possible for U or V statistics which consist of one sample. Our proof strategy for the SEP and UC results can be used generally in other contexts where two-sample problems arise. Given the SEP of relevant objects, asymptotic theory in various econometrics problems can be relatively easily developed (c.f. Andrews, 1994). In this respect, we can say that our results and proof extend applicability of the stochastic equicontinuity based technique to two-sample cases.

We note that derivations of asymptotic normality of estimators usually require the consistency of the estimators. This is also the case here, and we present the consistency of $(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}})$ in Lemmas 1 and 2. The consistency result of $\hat{\boldsymbol{\alpha}}$ is quite standard ($\hat{\boldsymbol{\alpha}}$ is the vector of probit and linear-regression estimators). We note that $\hat{\theta}^\lambda$ is a two-stage estimator which depends upon the preliminary estimator $\hat{\boldsymbol{\alpha}}$ (while we formalize $(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}})$ as a one-stage moment estimator which simultaneously solves a system of moment equations). For the consistency of $\hat{\theta}^\lambda$, we apply the result from CLK, who analyzed two-stage estimators when objective functions are not smooth (recall again that the empirical objective function $M_n^\lambda(\cdot, \cdot)$ is not continuous). While CLK's focus seems to be the case when preliminary (first-stage) estimators are nonparametric, their consistency theorem is still applicable to the case when a preliminary estimator is parametric like ours. A key in applying CLK's consistency result is also the

SEP of $\left[M_n^{\lambda}(\theta,\boldsymbol{\alpha}) - \bar{M}^{\lambda}(\theta,\boldsymbol{\alpha})\right] = \nu_n(\theta,\boldsymbol{\alpha})/\sqrt{n}$, which follows from that of $\nu_n(\theta,\boldsymbol{\alpha})$.

Conditions for deriving the asymptotic distribution of $\hat{\theta}^{\lambda}(g,h)$: We here present a set of conditions required for our asymptotic result. First, we work with the following asymptotic scheme:

Assumption 1 *Let N and n be the numbers of observations of past and current, respectively; N_g and n_g be those of g-type observations of past and current data (N_h and n_h are defined analogously). There exist some constants $c_g^P, c_g \in (0,1)$ and $\bar{c} \in (0,\infty)$, such that $N_g/N \to c_g^P$ as $N_g, N \to \infty$, and $n_g/n \to c_g$ as $n_g, n \to \infty$ ($N = N_g + N_h$; $n = n_g + n_h$), and $n/N \to \bar{c}$ as $n, N \to \infty$.*

This implies that the ratios of g-type applicants do not degenerate in both the past and current data (as $n, N \to \infty$), which also means that the limit ratios of h-type applicants are well-defined $c_h^P = 1 - c_g^P$ and $c_h := 1 - c_g$. In our subsequent asymptotic analysis, we suppose "$n, n_g \to \infty$" and/or "$N, N_g \to \infty$", which we often denote only by "$n \to \infty$" for notational simplicity.

Assumption 2 *(i) There exist some compact interval Θ^{λ} in $\mathbb{R}$ and some compact set $\mathcal{A}$ in $\mathbb{R}^{4d}$ such that $\theta_0^{\lambda} \in \mathrm{Int}\left(\Theta^{\lambda}\right)$ and $\alpha_0 \in \mathrm{Int}(\mathcal{A})$. (ii) The matrix $E[X_i^P (X_i^P)' | G_i^P]$ is invertible given any realization of G_i^P; $E[X_i X_i' | G_i]$ is also invertible given any realization of G_i. (iii) $\left\{(Y_i^P, X_i^P, G_i^P)\right\}_{i=1}^N$ and $\{(X_i, G_i)\}_{i=1}^n$ are I.I.D. sequences of random vectors, and they are mutually independent. (iv) The supports of (Y_i^P, X_i^P, G_i^P) and (X_i, G_i) are bounded.*

Assumption 3 *(i) Each component of X_i is either continuously-distributed with support which is some compact subset of $\mathbb{R}$ or discretely distributed with support of finite elements (i.e., there is no component whose distribution is a mixture of continuous and discrete ones). At least one component of $X_i = (X_{i,1}, \ldots, X_{i,d})'$ is continuously distributed conditionally on G_i and the other components of X_i, and each of coefficients of $\boldsymbol{\alpha}_0$ associated to the continuously-distributed components is not zero.*
(ii) Let $X_{i,1}$ be the first component of X_i and the continuously-distributed component given in the part i). There exists the conditional probability density of $X_{i,1}$, $f_g(x_1 | x_2, \ldots, x_d)$, given $G_i = g$ and the other components $X_{i,-1}(:= (X_{i,2}, \ldots, X_{i,d})') = x_{-1}(:= (x_2, \ldots, x_d)')$, that is, f_g satisfying

$$\int_{A_1}\int_{A_{-1}} f_g(x_1|x_{-1})\, dx_1 \Pr\left[dx_{-1} \in A_{-1}, G = g\right] = \Pr\left[X_{i,1} \in A_1, X_{i,-1} \in A_{-1}, G_i = g\right],$$

for any g (in the support of G_i), and for any Borel sets A_1 and A_{-1} (on the supports of $X_{i,1}$ and $X_{i,-1}$, respectively).
(iii) For any g (in the support of G_i),

$$\sup f_g(x_1|x_{-1}) < \infty,$$

where the supremum is taken over the support of X_i. For any g (in the supports of G_i), $f_g(x_1|x_{-1})$ is continuous in the continuously-distributed components of $x = (x_1, \ldots, x_d)'$, and for any x_{-1}, $f_g(x_1|x_{-1})$ is continuously differentiable in x_1 on the closure of $\{z \in \mathbb{R} \mid f_g(z|x_{-1}) > 0\}$, and the first derivative satisfies

$$\sup f_g'(x_1|x_{-1}) < \infty,$$

where the supremum is taken over the support of X_i.

Assumption 4 *The probability density $f_{I_g - I_h}(\tau)$ of $I_{g,i} - I_{h,j} := X_{g,i}'\boldsymbol{\beta}_{0,g} - X_{h,j}'\boldsymbol{\beta}_{0,h}$ is strictly positive when $E[\mathbf{1}\{X_{g,i}x \succeq_\varepsilon X_{h,j}, X_{g,i}'\boldsymbol{\delta}_g \leq X_{h,j}'\boldsymbol{\delta}_h\}|X_{g,i}'\boldsymbol{\beta}_g - X_{h,j}'\boldsymbol{\beta}_h = \tau] > 0$.*

The conditions of Assumption 2 are quite standard. Recall the notations of $X_{g,i}$ and $X_{h,j}$ (introduced in Section 5). Then, by (iii), we can see that $X_{g,i}$ and $X_{h,j}$ are independent. Assumption 3, together with additive separability of the indices ($x_g'\boldsymbol{\beta}_g$, $x_h'\boldsymbol{\beta}_h$, $x_g'\boldsymbol{\delta}_g$ and $x_h'\boldsymbol{\delta}_h$), ensures the smoothness of $\bar{M}^\lambda(\theta, \boldsymbol{\alpha})$ with respect to θ and $\boldsymbol{\alpha}$. This can be verified by checking the smoothness of two components which constitute $\bar{M}^\lambda(\theta, \boldsymbol{\alpha})$, $E[\mathbf{1}\{X_{g,i}'\boldsymbol{\delta}_g \leq X_{h,j}'\boldsymbol{\delta}_h, X_{g,i}x \succeq_\varepsilon X_{h,j}\}]$ and $E[\mathbf{1}\{X_{g,i}'\boldsymbol{\beta}_g - X_{h,j}'\boldsymbol{\beta}_h \leq \theta\} \times \mathbf{1}\{X_{g,i}'\boldsymbol{\delta}_g \leq X_{h,j}'\boldsymbol{\delta}_h, X_{g,i}x \succeq_\varepsilon X_{h,j}\}]$. We can show the former is partially continuously differentiability in $(\boldsymbol{\delta}_g, \boldsymbol{\delta}_h)$, and the latter is so in $(\theta, \boldsymbol{\alpha})$, whose proof is omitted for brevity. Another implication of Assumption 3 is that the index variables $J_{g,i} = X_{g,i}'\boldsymbol{\delta}_{0,g}$ and $J_{h,j} = \tilde{X}_{h,j}'\boldsymbol{\delta}_{0,h}$ have their conditional probability densities given $X_{g,i,-1} = x_{-1}$ and $X_{h,j,-1} = \tilde{x}_{-1}$, respectively, i.e., $f_{J_g|X_{g,-1}}(s|x_{-1})$ and $f_{J_h|X_{h,-1}}(s|\tilde{x}_{-1})$. And these $f_{J_g|X_{g,-1}}(s|x_{-1})$ and $f_{J_h|X_{h,-1}}(\tilde{s}|\tilde{x}_{-1})$ satisfy the continuous differentiability (in s and $\tilde{s}$, respectively), as well as the uniform boundedness of themselves and their derivatives. Assumption 4 guarantees the identification of the quantile-based lower bound θ_0^λ – see the proof of Lemma 2.

Asymptotic distribution theorem: Given the above conditions, we can now state our distribution result:

Theorem 1 *Suppose that Assumptions 1, 2, 3 and 4 hold. Then,*

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \hat{\theta}^\lambda - \theta_0^\lambda \end{pmatrix} \xrightarrow{d} \begin{bmatrix} -(\partial/\partial\boldsymbol{\alpha}')\bar{\mathbf{L}}(\boldsymbol{\alpha}_0) & 0 \\ (\partial/\partial\boldsymbol{\alpha}')\bar{M}^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0) & (\partial/\partial\theta)\bar{M}^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0) \end{bmatrix}^{-1} \times N\left(0, \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{L}} & \boldsymbol{\Omega}_{\mathbf{L},M}^\lambda \\ \cdot & \Omega_M^\lambda \end{bmatrix}\right). \tag{D.3}$$

for each $\lambda \in (0,1)$, and for any (g,h) with $g \neq h$, where the forms of the asymptotic variance components, $\boldsymbol{\Omega}_{\mathbf{L}}$, $\boldsymbol{\Omega}_{\mathbf{L},M}^\lambda$ and Ω_M^λ, are given in the proof.

This (D.3) allows us to obtain the asymptotic distribution of $\sqrt{n}[\hat{\theta}^{\lambda} - \theta_0^{\lambda}]$. We note that the derivative components of the limit distribution can be computed in a straightforward manner: $(\partial/\partial\boldsymbol{\alpha}')\,\bar{\mathbf{L}}(\boldsymbol{\alpha}_0)$ is the derivative of the moment defined in (D.7) for the linear and probit estimators; $(\partial/\partial\boldsymbol{\alpha}')\,\bar{M}^{\lambda}\left(\theta_0^{\lambda}, \boldsymbol{\alpha}_0\right)$ and $(\partial/\partial\theta)\,\bar{M}^{\lambda}\left(\theta_0^{\lambda}, \boldsymbol{\alpha}_0\right)$ are computed based on the moment (D.9), where we note that M_n^{λ} is not differentiable but its limit $\bar{M}^{\lambda}$ is so under Assumption 3. $(\partial/\partial\boldsymbol{\alpha}')\,\bar{\mathbf{L}}(\boldsymbol{\alpha}_0)$ can be consistently estimated in a standard manner, while derivatives of $\bar{M}^{\lambda}\left(\theta_0^{\lambda}, \boldsymbol{\alpha}_0\right)$ can be estimated by using a kernel-based method. Note that if $\boldsymbol{\alpha}_0$ is known, we have $\sqrt{n}[\hat{\theta}^{\lambda} - \theta_0^{\lambda}] \xrightarrow{d} [(\partial/\partial\theta)\,\bar{M}^{\lambda}\left(\theta_0^{\lambda}, \boldsymbol{\alpha}_0\right)]^{-1} N\left(0, \Omega_M^{\lambda}\right)$, and we can see that the limit distribution in (D.3) allows us to evaluate estimation errors due to the preliminary step of the estimation of $\boldsymbol{\alpha}_0$.

The proof of Theorem 1: We first re-define the first-step estimator $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\beta}}_g', \hat{\boldsymbol{\beta}}_h', \hat{\boldsymbol{\delta}}_g', \hat{\boldsymbol{\delta}}_h')'$ as a moment-based one. Let

$$\begin{aligned}
&\mathbf{R}_{g,i}^P(\boldsymbol{\beta}_g) := X_{g,i}^P[Y_{g,i}^P - (X_{g,i}^P)'\boldsymbol{\beta}_g];\\
&\mathbf{S}_{g,i}(\boldsymbol{\delta}_g) := X_{g,i}\left[\frac{D_{g,i}\phi(X_{g,i}'\boldsymbol{\delta}_g)}{\Phi(X_{g,i}'\boldsymbol{\delta}_g)} - \frac{(1-D_{g,i})\Phi(X_{g,i}'\boldsymbol{\delta}_g)}{1-\Phi(X_{g,i}'\boldsymbol{\delta}_g)}\right];\\
&\mathbf{R}_{h,j}^P(\boldsymbol{\beta}_h) := X_{h,j}^P[Y_{h,j}^P - (X_{h,j}^P)'\boldsymbol{\beta}_h];\\
&\mathbf{S}_{h,j}(\boldsymbol{\delta}_h) := X_{h,j}\left[\frac{D_{h,j}\phi(X_{h,j}\boldsymbol{\delta}_h)}{\Phi(X_{h,j}'\boldsymbol{\delta}_h)} - \frac{(1-D_{h,j})\phi(X_{h,j}'\boldsymbol{\delta}_h)}{1-\Phi(X_{h,j}'\boldsymbol{\delta}_h)}\right],
\end{aligned} \tag{D.4}$$

where ϕ and Φ are the density and distribution functions of the standard normal, respectively, and we note that these four components are mutually independent (under Assumption 2). Then, we let the (pseudo) true parameter $\boldsymbol{\alpha}_0 = (\boldsymbol{\beta}_{0,g}', \boldsymbol{\beta}_{0,h}', \boldsymbol{\delta}_{0,g}', \boldsymbol{\delta}_{0,h}')'$ as the solution to the following moment equations:

$$\bar{\mathbf{L}}(\boldsymbol{\alpha}) = 0, \tag{D.5}$$

and define $\hat{\boldsymbol{\alpha}}$ as the solution to its sample counterpart:

$$\hat{\mathbf{L}}(\boldsymbol{\alpha}) = 0. \tag{D.6}$$

where $\bar{\mathbf{L}}(\boldsymbol{\alpha})$ and $\mathbf{L}(\boldsymbol{\alpha})$ are defined, corresponding to the linear and probit models, as follows:

$$\bar{\mathbf{L}}(\boldsymbol{\alpha}) := \begin{pmatrix} E[\mathbf{R}_{g,i}^P(\boldsymbol{\beta}_g)] \\ E[\mathbf{R}_{h,j}^P(\boldsymbol{\beta}_h)] \\ E\left[\mathbf{S}_{g,i}(\boldsymbol{\delta}_g)\right] \\ E[\mathbf{S}_{h,j}(\boldsymbol{\delta}_h)] \end{pmatrix}; \quad \text{and} \quad \mathbf{L}_n(\boldsymbol{\alpha}) := \begin{pmatrix} N_g^{-1}\sum_{i=1}^{N_g}\mathbf{R}_{g,i}^P(\boldsymbol{\beta}_g) \\ N_h^{-1}\sum_{j=1}^{N_h}\mathbf{R}_{h,j}^P(\boldsymbol{\beta}_h) \\ n_g^{-1}\sum_{i=1}^{n_g}\mathbf{S}_{g,i}(\boldsymbol{\delta}_g) \\ n_h^{-1}\sum_{j=1}^{n_h}\mathbf{S}_{h,j}(\boldsymbol{\delta}_h) \end{pmatrix}. \tag{D.7}$$

By this definition of $\hat{\boldsymbol{\alpha}}$ as a moment-based estimator, the asymptotic distribution result derived below is valid irrespective of the correct specification of the linear and/or probit models (in particular, the probit estimator $(\hat{\boldsymbol{\delta}}_g, \hat{\boldsymbol{\delta}}_h)$ should be interpreted as a quasi maximum likelihood estimator, whose asymptotic variance is given by a so-called *sandwich* form).

Before deriving the asymptotic distribution result of $\hat{\theta}^{\lambda}$, we provide two lemmas on the consistency:

Lemma 1 *Suppose that Assumptions 1, 2 and 3 hold. Then,* $\hat{\boldsymbol{\alpha}}$ *is consistent, i.e.,* $||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0|| \xrightarrow{P} 0$ *as* $n \to \infty$.

Proof. The consistency of $\hat{\boldsymbol{\alpha}}$ follows from quite standard arguments for parametric estimation, e.g., Section 2 of of Newey and McFadden (1994): the parameter space $\mathcal{A}$ is supposed to be compact in Assumption 2; the identification and the uniform convergence of $\mathbf{L}_n$ can be verified by using the invertibility and I.I.D. conditions in Assumption 2, the linear and probit specifications, and the boundedness of relevant functions and variables. ■

Lemma 2 *Suppose that Assumptions 1, 2, 3 and 4 hold. Then,* $\hat{\theta}^{\lambda}$ *is consistent, i.e.,* $|\hat{\theta}^{\lambda} - \theta_0^{\lambda}| \xrightarrow{P} 0$ *as* $n \to \infty$.

Proof. We can verify the consistency of $\hat{\theta}^{\lambda}$ by using, e.g., Theorem 1 of CLK (2003). To check the identification of θ_0^{λ}, recall that θ_0^{λ} is defined as a parameter satisfying $\bar{M}^{\lambda}\left(\theta_0^{\lambda}, \boldsymbol{\alpha}_0\right) = 0$. Since we can write

$$\begin{aligned} &E[\mathbf{1}\{X'_{g,i}\boldsymbol{\beta}_g - X'_{h,j}\boldsymbol{\beta}_h \leq \theta\} \times \mathbf{1}\{X_{g,i}x \succeq_{\varepsilon} X_{h,j}, X'_{g,i}\boldsymbol{\delta}_g \leq X'_{h,j}\boldsymbol{\delta}_h\}] \\ &= \int_{-\infty}^{\theta} E[\mathbf{1}\{X_{g,i} \succeq_{\varepsilon} X_{h,j}, X'_{g,i}\boldsymbol{\delta}_g \leq X'_{h,j}\boldsymbol{\delta}_h\} | X'_{g,i}\boldsymbol{\beta}_g - X'_{h,j}\boldsymbol{\beta}_h = \tau] f_{\tau}(t)\, dt, \end{aligned}$$

and this is strictly increasing in θ by Assumption 4, $\left\{\theta : \bar{M}^{\lambda}(\theta, \boldsymbol{\alpha}_0) = 0\right\}$ consists of a single point. Therefore, by the continuity of $\bar{M}^{\lambda}(\cdot, \boldsymbol{\alpha}_0)$ (implied by Assumption 3), and the compactness of the parameter space Θ (imposed in Assumption 2), the identification condition of Theorem 1 is satisfied. Given the identification condition, the consistency of $\hat{\boldsymbol{\alpha}}$, the compactness of Θ , the continuity of the limit function $\bar{M}^{\lambda}(\cdot, \cdot)$ (implied by Assumption 3), the stochastic equicontinuity M_n^{λ} (verified in Lemma 3), we can check all the conditions of Theorem 1 of CLK, and the desired result follows. ■

Now, define the following objects:

$$\begin{aligned} M_n^{\lambda}(\theta, \boldsymbol{\alpha}) \quad &: \quad = \frac{1}{n_g n_h} \sum\nolimits_{i=1}^{n_g} \sum\nolimits_{j=1}^{n_h} \psi^{\lambda}(X_{g,i}, X_{h,j}, \theta, \boldsymbol{\alpha}) \\ &= \quad \int\int \psi^{\lambda}(x, \tilde{x}, \theta, \boldsymbol{\alpha})\, d\hat{F}_h(\tilde{x})\, d\hat{F}_g(x) \end{aligned} \tag{D.8}$$

and

$$\bar{M}^{\lambda}(\theta, \boldsymbol{\alpha}) := \int\int \psi^{\lambda}(x, \tilde{x}, \theta, \boldsymbol{\alpha})\, dF_h(\tilde{x})\, dF_g(x), \tag{D.9}$$

where

$$\psi^{\lambda}(x, \tilde{x}, \theta, \boldsymbol{\alpha}) := \left[\lambda - \mathbf{1}\{x'\boldsymbol{\beta}_g - \tilde{x}'\boldsymbol{\beta}_h \leq \theta\}\right] \times \mathbf{1}\left\{x'\boldsymbol{\delta}_g \leq \tilde{x}'\boldsymbol{\delta}_h,\ x \succeq_{\varepsilon} \tilde{x}\right\},$$

F_g and F_h are the conditional distribution functions of X_i, given $G_i = g$ and h, respectively, and $\hat{F}_g$ and $\hat{F}_h$ are their empirical distribution functions. This definition of M_n^λ coincides with the one in Section 5. Recalling the definitions of $\theta^\lambda(g,h)$, we write the (pseudo) true parameter $\theta_0^\lambda = \theta_0^\lambda(g,h)$ as the one satisfying $\bar{M}^\lambda\left(\theta_0^\lambda, \boldsymbol{\alpha}_0\right) = 0$, and its estimator $\hat{\theta}^\lambda = \hat{\theta}^\lambda(g,h)$ as the one satisfying $M_n(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}}) = 0$. Then, we can see the estimator $(\hat{\boldsymbol{\alpha}}', \hat{\theta}^\lambda)'$ satisfies

$$\begin{pmatrix} \mathbf{L}_n(\hat{\boldsymbol{\alpha}}) \\ M_n(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}}) \end{pmatrix} = 0. \tag{D.10}$$

By exploiting this expression as a solution to the simultaneous moment equations, we subsequently derive the asymptotic distribution of the estimator, quantifying the effect of the preliminary estimation of $\hat{\boldsymbol{\alpha}}$ on $\hat{\theta}^\lambda$ (c.f. Newey, 1984).

We now consider the asymptotic expansion of the LHS of (D.10). $\mathbf{L}_n(\hat{\boldsymbol{\alpha}})$ can be written as

$$0 = \sqrt{n}\mathbf{L}_n(\hat{\boldsymbol{\alpha}}) = \sqrt{n}\mathbf{L}_n(\boldsymbol{\alpha}_0) + (\partial/\partial\boldsymbol{\alpha}')\,\mathbf{L}_n(\tilde{\boldsymbol{\alpha}})\sqrt{n}\left[\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right],$$

where the differentiability of $\mathbf{L}_n$ follows from its linear and probit specifications, and each element of $\tilde{\boldsymbol{\alpha}}$ is on the line segment connecting a corresponding component of $\hat{\boldsymbol{\alpha}}$ to that of $\boldsymbol{\alpha}_0$. We note that the normalization factor is given by $\sqrt{n}$, which is justified by Assumption 1 but requires adjustments in the asymptotic variance by some constants such as $\bar{c}, c_g^P$ and c_g as we will see later. Given the results of Lemma 1, we have

$$(\partial/\partial\alpha')\,\mathbf{L}_n(\boldsymbol{\alpha}_0)\sqrt{n}\left[\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right]\left[1 + o_P(1)\right] = -\sqrt{n}\mathbf{L}_n(\boldsymbol{\alpha}_0). \tag{D.11}$$

To consider the expansion of $M_n(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}})$, we introduce the following objects:

$$\psi_1^\lambda(x,\theta,\boldsymbol{\alpha}) \;:=\; \int \psi^\lambda(x,\tilde{x},\theta,\boldsymbol{\alpha})\,dF_h(\tilde{x}); \quad \psi_2^\lambda(\tilde{x},\theta,\boldsymbol{\alpha}) := \int \psi^\lambda(x,\tilde{x},\theta,\boldsymbol{\alpha})\,dF_g(x) \tag{D.12}$$

$$M_{1,n}^\lambda(\theta,\boldsymbol{\alpha}) \;:=\; \frac{1}{n_g}\sum\nolimits_{j=1}^{n_g}\psi_1^\lambda(X_{g,i},\theta,\boldsymbol{\alpha}); \quad M_{2,n}^\lambda(\theta,\boldsymbol{\alpha}) := \frac{1}{n_h}\sum\nolimits_{j=1}^{n_h}\psi_2^\lambda(X_{h,j},\theta,\boldsymbol{\alpha}) \tag{D.13}$$

where we note that ψ_1^λ and ψ_2^λ are sort of projections objects (analogous objects often appear in considering U statistics). $M_{1,n}^\lambda$ and $M_{2,n}^\lambda$ satisfy

$$E[M_{1,n}^\lambda(\theta,\boldsymbol{\alpha})] = E[M_{2,n}^\lambda(\theta,\boldsymbol{\alpha})] = \bar{M}^\lambda(\theta,\boldsymbol{\alpha}) = 0 \;\text{ at } (\theta,\boldsymbol{\alpha}) = (\theta_0^\lambda, \boldsymbol{\alpha}_0).$$

We then obtain

$$\begin{aligned} 0 &= \sqrt{n}[M_n^\lambda(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}}) - \bar{M}^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0)] \\ &= \sqrt{n}[\bar{M}^\lambda(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}}) - \bar{M}^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0)] - \sqrt{n}[M_n^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0) - \bar{M}^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0)] \\ &\quad + \sqrt{n}[M_n^\lambda(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}}) - \bar{M}^\lambda(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}}) + M_n^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0) - \bar{M}^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0)]. \end{aligned} \tag{D.14}$$

By the continuous differentiability of $\bar{M}^\lambda$, the first term on the RHS can be further expanded as

$$\begin{aligned}
& \sqrt{n}[\bar{M}^\lambda(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}}) - \bar{M}^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0)] \\
= \; & (\partial/\partial\theta)\, \bar{M}^\lambda(\tilde{\theta}^\lambda, \tilde{\boldsymbol{\alpha}})\sqrt{n}[\hat{\theta}^\lambda - \theta_0^\lambda] + (\partial/\partial\boldsymbol{\alpha}')\, \bar{M}^\lambda(\tilde{\theta}^\lambda, \tilde{\boldsymbol{\alpha}})\sqrt{n}\,[\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0] \\
= \; & (\partial/\partial\theta)\, \bar{M}^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0)\sqrt{n}[\hat{\theta}^\lambda - \theta_0^\lambda] + (\partial/\partial\boldsymbol{\alpha}')\, \bar{M}^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0)\sqrt{n}\,[\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0] \\
& + o_P(1), \qquad \text{(D.15)}
\end{aligned}$$

where $\tilde{\theta}^\lambda$ and $\tilde{\boldsymbol{\alpha}}$ are on the line segment connecting $\hat{\theta}^\lambda$ to θ_0^λ and $\hat{\boldsymbol{\alpha}}$ to $\boldsymbol{\alpha}_0$, respectively; and the last equality follows from the continuity of the derivatives and the consistency of $(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}})$.

To investigate asymptotic behavior of the second and third terms on the RHS of (D.14), we consider $\left\{\nu_n^\lambda(\theta, \boldsymbol{\alpha})\right\}_{(\theta,\boldsymbol{\alpha})\in\Theta\times\mathcal{A}}$, a stochastic process indexed by $(\theta, \boldsymbol{\alpha})$:

$$\nu_n^\lambda(\theta, \boldsymbol{\alpha}) := \sqrt{n}[M_n^\lambda(\theta, \boldsymbol{\alpha}) - \bar{M}^\lambda(\theta, \boldsymbol{\alpha})]. \qquad \text{(D.16)}$$

This $\left\{\nu_n^\lambda(\theta, \boldsymbol{\alpha})\right\}$ has the following desirable properties:

Lemma 3 *Suppose that Assumptions 1, 2, and 3 hold. Then, (i) it holds that as $n \to \infty$,*

$$\begin{aligned}
\nu_n^\lambda(\theta, \boldsymbol{\alpha}) \;=\; & \sqrt{n}[M_{1,n}^\lambda(\theta, \boldsymbol{\alpha}) - \bar{M}^\lambda(\theta, \boldsymbol{\alpha})] + \sqrt{n}[M_{2,n}^\lambda(\theta, \boldsymbol{\alpha}) - \bar{M}^\lambda(\theta, \boldsymbol{\alpha})] \\
& + O_P(\sqrt{(\log n)/n}), \qquad \text{(D.17)}
\end{aligned}$$

uniformly over $(\theta, \boldsymbol{\alpha}) \in \Theta^\lambda \times \mathcal{A}$. (ii) The stochastic process $\left\{\nu_n^\lambda(\theta, \boldsymbol{\alpha})\right\}_{(\theta,\boldsymbol{\alpha})\in\Theta^\lambda\times\mathcal{A}}$ is stochastically equicontinuous with respect to the pseudo metric:

$$\begin{aligned}
\rho\left((\theta_1, \boldsymbol{\alpha}_1), (\theta_2, \boldsymbol{\alpha}_2)\right) \;:\; = & \left\{\int \left|\psi_1^\lambda(x, \theta_1, \boldsymbol{\alpha}_1) - \psi_1^\lambda(x, \theta_2, \boldsymbol{\alpha}_2)\right|^2 dF_g(x)\right\}^{1/2} \\
& + \left\{\int \left|\psi_2^\lambda(\tilde{x}, \theta_1, \boldsymbol{\alpha}_1) - \psi_2^\lambda(\tilde{x}, \theta_2, \boldsymbol{\alpha}_2)\right|^2 dF_h(\tilde{x})\right\}^{1/2}.
\end{aligned}$$

The proof of the lemma is provided below.

Remark 1 *As we subsequently see, the first result of Lemma 3 immediately implies the asymptotic normality of $\nu_n^\lambda\left(\theta_0^\lambda, \boldsymbol{\alpha}_0\right) = \sqrt{n}M_n^\lambda\left(\theta_0^\lambda, \boldsymbol{\alpha}_0\right)$ ($\bar{M}^\lambda\left(\theta_0^\lambda, \boldsymbol{\alpha}_0\right) = 0$), since projections objects $M_{1,n}^\lambda\left(\theta_0^\lambda, \boldsymbol{\alpha}_0\right)$ and $M_{2,n}^\lambda(\theta, \boldsymbol{\alpha})$ are independent and each of them is computed based on a single summation, to which standard CLT results are applicable.*

Using ν_n^λ in (D.16), we can write the third term on the RHS of (D.14) as

$$\sqrt{n}[M_n^\lambda(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}}) - \bar{M}^\lambda(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}}) + M_n^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0) - \bar{M}^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0)] = \nu_n^\lambda(\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}}) - \nu_n^\lambda(\theta_0^\lambda, \boldsymbol{\alpha}_0). \qquad \text{(D.18)}$$

For notational simplicity, write $\hat{\boldsymbol{\vartheta}} := (\hat{\theta}^\lambda, \hat{\boldsymbol{\alpha}})$ and $\boldsymbol{\vartheta}_0 := (\theta_0^\lambda, \boldsymbol{\alpha}_0)$ for now. Then, given any $a, \epsilon > 0$, there exists some $b > 0$ such that

$$\begin{aligned}
& \limsup_{n\to\infty} \Pr[|\nu_n^\lambda(\hat{\boldsymbol{\vartheta}}) - \nu_n^\lambda(\boldsymbol{\vartheta}_0)| > a] \\
\leq \; & \limsup_{n\to\infty} \Pr[|\nu_n^\lambda(\hat{\boldsymbol{\vartheta}}) - \nu_n^\lambda(\boldsymbol{\vartheta}_0)| > a,\ \rho(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta}_0) \leq b] + \limsup_{n\to\infty} \Pr[\rho(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta}_0) > b] \\
\leq \; & \limsup_{n\to\infty} \Pr[\sup_{\rho(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta}_0) < b} |\nu_n^\lambda(\hat{\boldsymbol{\vartheta}}) - \nu_n^\lambda(\boldsymbol{\vartheta}_0)| > a] < \epsilon,
\end{aligned}$$

where the second inequality use the consistency of $\hat{\boldsymbol{\vartheta}} = (\hat{\theta}^{\lambda}, \hat{\boldsymbol{\alpha}})$ (Lemmas 1 and 2), the continuity of ρ (implied by Assumption 3), and the continuous mapping $(\rho(\hat{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta}_0) \xrightarrow{P} 0)$; and the last inequality holds by (ii) of Lemma 3. Since a and ϵ can be arbitrary, it holds that

$$|\nu_n^{\lambda}(\hat{\theta}^{\lambda}, \hat{\boldsymbol{\alpha}}) - \nu_n^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0)| = o_P(1). \tag{D.19}$$

By (D.14), (D.15), and (D.18)-(D.19) with noting that $\bar{M}^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0) = 0$, we now obtain

$$(\partial/\partial\theta)\,\bar{M}^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0)\sqrt{n}[\hat{\theta}^{\lambda} - \theta_0^{\lambda}] + (\partial/\partial\boldsymbol{\alpha}')\bar{M}^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0)\sqrt{n}\,[\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0] = \sqrt{n}M_n^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0) + o_P(1). \tag{D.20}$$

Putting (D.11) and (D.20) together, we can obtain

$$\begin{aligned}\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \hat{\theta}^{\lambda} - \theta_0^{\lambda} \end{pmatrix} &= \begin{bmatrix} -(\partial/\partial\boldsymbol{\alpha}')\,\mathbf{L}_n(\boldsymbol{\alpha}_0) & 0 \\ (\partial/\partial\boldsymbol{\alpha}')\,\bar{M}^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0) & (\partial/\partial\theta)\,\bar{M}^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0) \end{bmatrix}^{-1} \\ &\quad \times\sqrt{n}\begin{pmatrix} \mathbf{L}_n(\boldsymbol{\alpha}_0) \\ M_n^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0) \end{pmatrix} + o_P(1). \end{aligned} \tag{D.21}$$

Now, by (D.17) with noting that $\bar{M}^{\lambda}\left(\theta_0^{\lambda}, \boldsymbol{\alpha}_0\right) = 0$, we can see that

$$\sqrt{n}M_n^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0) = \sqrt{n}M_{1,n}^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0) + \sqrt{n}M_{2,n}^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0) + o_P(1). \tag{D.22}$$

By this expression, the CLT for I.I.D. sequences, the independence between $M_{1,n}^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0)(= n_g^{-1}\sum_{i=1}^{n_g}\psi_1(X_{g,i}, \theta_0, \boldsymbol{\alpha}_0))$ and $M_2^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0)(= n_h^{-1}\sum_{j=1}^{n_h}\psi_2(X_{h,j}, \theta_0, \boldsymbol{\alpha}_0))$, and Assumption 1, we have

$$\sqrt{n}[M_n^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0) - \bar{M}^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0)] \xrightarrow{d} N(0, \Omega_M^{\lambda}), \tag{D.23}$$

where

$$\Omega_M^{\lambda} := E[c_g^{-1}|\psi_1^{\lambda}(X_{g,i}, \theta_0^{\lambda}, \boldsymbol{\alpha}_0)|^2 + c_h^{-1}|\psi_2^{\lambda}(X_{h,j}, \theta_0^{\lambda}, \boldsymbol{\alpha}_0)|^2].$$

This form of Ω_M^{λ} is computed as the sum of variances of the "projection" components, which is similar to asymptotic variances of U statistics, where adjustments by c_g and c_h are required since the normalization factor is given by $\sqrt{n}$.

By the definition of $\mathbf{L}_n(\boldsymbol{\alpha}_0)$ given in (D.7), the expression of $\sqrt{n}M_n^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0)$ in (D.22), and the boundedness of relevant components, and the CLT for I.I.D. sequences with the aid of the Cramér-Wold device, we can obtain the joint CLT result:

$$\sqrt{n}\begin{pmatrix} \mathbf{L}_n(\boldsymbol{\alpha}_0) \\ M_n^{\lambda}(\theta_0^{\lambda}, \boldsymbol{\alpha}_0) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \boldsymbol{\Omega}_{\mathbf{L}} & \boldsymbol{\Omega}_{\mathbf{L},M}^{\lambda} \\ \cdot & \Omega_M^{\lambda} \end{pmatrix}\right). \tag{D.24}$$

where the variance components can be written as

$$\begin{aligned} \boldsymbol{\Omega}_{\mathbf{L}} &: = E\left[\mathrm{diag}\left\{\left(\bar{c}/c_g^P\right)\mathbf{R}_{0,g,i}^P(\mathbf{R}_{0,g,i}^P)',\ \left(\bar{c}/c_h^P\right)\mathbf{R}_{0,h,j}^P(\mathbf{R}_{0,h,j}^P)',\ c_g^{-1}\mathbf{S}_{0,g,i}\mathbf{S}_{0,g,i}',\ c_h^{-1}\mathbf{S}_{0,h,j}\mathbf{S}_{0,h,j}'\right\}\right]; \\ \boldsymbol{\Omega}_{\mathbf{L},M}^{\lambda} &: = E\left[(\mathbf{0}',\ c_g^{-1}\psi_1^{\lambda}(X_{g,i}, \theta_0^{\lambda}, \boldsymbol{\alpha}_0)\mathbf{S}_{0,g,i}',\ c_h^{-1}\psi_2^{\lambda}(X_{h,j}, \theta_0^{\lambda}, \boldsymbol{\alpha}_0)\mathbf{S}_{0,h,j}')'\right]; \end{aligned}$$

where diag $\{A, B, \dots\}$ stands for a block-diagonal matrix with square matrices $A, B, \dots$ as block-diagonal elements; $\mathbf{R}^P_{0,g,i}$, $\mathbf{R}^P_{0,h,j}$, $\mathbf{S}_{0,g,i}$ and $\mathbf{S}_{0,h,j}$ denote $\mathbf{R}^P_{g,i}(\boldsymbol{\beta}_{0,g})$, $\mathbf{R}^P_{h,j}(\boldsymbol{\beta}_{0,h})$, $\mathbf{S}_{g,i}(\boldsymbol{\delta}_{0,g})$ and $\mathbf{S}_{h,j}(\boldsymbol{\delta}_{0,h})$ (evaluated at the pseudo true values; defined in (D.4), respectively; $\boldsymbol{\Omega}_{\mathbf{L}}$ is a $4d$-by-$4d$ matrix; $\boldsymbol{\Omega}^{\lambda}_{\mathbf{L},M}$ is a $4d$-by-1 matrix ($\mathbf{0}$ is a $2d$-by-1 vector of zeros); and Ω^{λ}_{M} is a scalar. The block-diagonal form of $\boldsymbol{\Omega}_{\mathbf{L}}$ is due to the independence between past and current cohorts' observations, and between g and h observations. Note again that coefficients $\bar{c}$, c^P_g, c^P_h, c_g and c_h (defined in Assumption 1) are required for adjusting the difference in the numbers of observations.

By the asymptotic expression (D.21) and the CLT result (D.24), we can obtain the desired result (D.3) of the theorem.

Proof of Lemma 3: Look at the decomposition:

$$\begin{aligned}\nu^{\lambda}_n(\theta,\boldsymbol{\alpha}) &= \sqrt{n}[M^{\lambda}_{1,n}(\theta,\boldsymbol{\alpha}) - \bar{M}^{\lambda}(\theta,\boldsymbol{\alpha})] + \sqrt{n}[M^{\lambda}_{2,n}(\theta,\boldsymbol{\alpha}) - \bar{M}^{\lambda}(\theta,\boldsymbol{\alpha})] \\ &\quad + \sqrt{n}[M^{\lambda}_{n}(\theta,\boldsymbol{\alpha}) - M^{\lambda}_{1,n}(\theta,\boldsymbol{\alpha}) - M^{\lambda}_{2,n}(\theta,\boldsymbol{\alpha}) + \bar{M}^{\lambda}(\theta,\boldsymbol{\alpha})] \\ &= : J^{\lambda}_n(\theta,\boldsymbol{\alpha}) + T^{\lambda}_n(\theta,\boldsymbol{\alpha}).\end{aligned}$$

We below verify the uniform negligibility of $T^{\lambda}_n(\theta,\boldsymbol{\alpha})$, i.e.,

$$|T^{\lambda}_n(\theta,\boldsymbol{\alpha})| = O_P(\sqrt{(\log n)/n}) \text{ uniformly over } (\theta,\boldsymbol{\alpha}) \in \Theta^{\lambda}\times\mathcal{A}, \tag{D.25}$$

and the stochastic equicontinuity of $\left\{J^{\lambda}_n(\theta,\boldsymbol{\alpha})\right\}_{(\theta,\boldsymbol{\alpha})\in\Theta^{\lambda}\times\mathcal{A}}$, a stochastic process indexed by $(\theta,\boldsymbol{\alpha}) \in \Theta^{\lambda}\times\mathcal{A}(\subset \mathbb{R}\times\mathbb{R}^{4d})$, with respect to the pseudo metric ρ, i.e., $\forall a, \epsilon > 0$, $\exists b > 0$ such that

$$\limsup_{n\to\infty} \Pr\left[\sup_{(\theta_1,\boldsymbol{\alpha}_1),(\theta_2,\boldsymbol{\alpha}_2)\in\Theta^{\lambda}\times\mathcal{A};\ \rho((\theta_1,\boldsymbol{\alpha}_1),(\theta_2,\boldsymbol{\alpha}_2))<b} |J^{\lambda}_n(\theta,\boldsymbol{\alpha})| > a\right] < \epsilon. \tag{D.26}$$

Then, we can immediately see that the conclusion (i) of the theorem follows from (D.25), and (ii) follows from (D.25) and (D.26).

To show (D.25) and (D.26), observe that components of $\nu^{\lambda}_n(\theta,\alpha)$ are computed based on

$$\psi^{\lambda}(x,\tilde{x},\theta,\boldsymbol{\alpha}) = \left[\lambda - \mathbf{1}\{x'\boldsymbol{\beta}_g - \tilde{x}'\boldsymbol{\beta}_h \le \theta\}\right] \times \mathbf{1}\{x'\boldsymbol{\delta}_g \le \tilde{x}'\boldsymbol{\delta}_h\}\,\mathbf{1}\{x \succeq_{\varepsilon} \tilde{x}\}. \tag{D.27}$$

We can think of

$$\mathcal{F}^{\lambda} := \left\{\psi^{\lambda}(x,\tilde{x},\theta,\alpha) \mid (\theta,\alpha) \in \Theta^{\lambda}\times\mathcal{A}\right\}$$

as a set of functions: $(x,\tilde{x})\,(\in \mathcal{X}_g\times\mathcal{X}_h \subset \mathbb{R}^{2d}\times\mathbb{R}^{2d}) \to \psi^{\lambda}(\in\mathbb{R})$ indexed by $(\theta,\alpha)\in\Theta^{\lambda}\times\mathcal{A}$.

We derive an upper bound of the uniform covering number of this $\mathcal{F}^{\lambda}$. To this end, look at the following two sets of functions: $\mathcal{F}^{\lambda}_1 := \{\lambda - \mathbf{1}\{x'\boldsymbol{\beta}_g - \tilde{x}'\boldsymbol{\beta}_h \le \theta\} \mid (\theta,\boldsymbol{\beta}_g,\boldsymbol{\beta}_h) \in \mathbb{R}\times\mathbb{R}^{2d}\times\mathbb{R}^{2d}\}$ and $\mathcal{F}_2 := \{\mathbf{1}\{x'\boldsymbol{\delta}_g - \tilde{x}'\boldsymbol{\delta}_h \le 0\} \mid (\boldsymbol{\delta}_g,\boldsymbol{\delta}_h) \in \mathbb{R}^{2d}\times\mathbb{R}^{2d}\}$. By Theorem 9.2,

Lemma 9.8, (iv)-(v) of Lemma 9.9, and (i) of Lemma 9.12 of Kosorok (2008), the L_r uniform covering numbers of $\mathcal{F}_1^\lambda$ and $\mathcal{F}_2$ are given by

$$\sup_Q N(2\varepsilon, \mathcal{F}_1^\lambda, L_r(Q)) \leq \Lambda_1 \varepsilon^{-r(4d+1)} \quad \text{and} \quad \sup_Q N(\varepsilon, \mathcal{F}_2, L_r(Q)) \leq \Lambda_2 \varepsilon^{-r(4d+1)}, \tag{D.28}$$

for $\varepsilon \in (0,1)$, where the supremum is taken over all probability measures on $\mathbb{R}^{2d} \times \mathbb{R}^{2d}$; and $\Lambda_1, \Lambda_2 (> 0)$ are some constants independent of Q (i.e., $\mathcal{F}_1^\lambda$ and $\mathcal{F}_2$ are VC classes). By the inclusion relation $\mathcal{F}^\lambda \subset \{f_1 \times f_2 | f_1 \in \mathcal{F}_1^\lambda,\ f_2 \in \mathcal{F}_2\}$, as well as by Lemma 2.14 of Pakes and Pollard (1989), the L_1 uniform covering number of $\mathcal{F}^\lambda$ satisfies

$$\sup_Q N(2\varepsilon, \mathcal{F}^\lambda, L_1(Q)) \leq A\varepsilon^{-V}, \tag{D.29}$$

for some constants $A, V \in (0, \infty)$ which are independent of Q (i.e., $\mathcal{F}_1^\lambda$ and $\mathcal{F}_2$ are Euclidean, and thus $\mathcal{F}^\lambda$ are so). By this bound of the uniform covering number, we can construct a collections of balls to cover $\mathcal{F}^\lambda$, $\{\mathcal{F}_k\}_{k=1}^\nu$ (for each Q) such that any element of $\mathcal{F}^\lambda$ is included in at least one of the balls, each $\mathcal{F}_k$ has its center ψ_k with $\int |\psi - \psi_k| \, dQ \leq 2\varepsilon$ for any $\psi \in \mathcal{F}_k$ and with $\nu \leq A\varepsilon^{-V}$. The center ψ_k is not necessarily contained in $\mathcal{F}^\lambda$ (see, e.g., page 18 of Kosorok, 2008), but we can pick some $\tilde{\psi}_k \in \mathcal{F}_k \cap \mathcal{F}^\lambda$ such that for any $\psi \in \mathcal{F}_k$,

$$\int |\psi - \tilde{\psi}_k| dQ \leq 4\varepsilon, \tag{D.30}$$

for each k (the existence of such $\tilde{\psi}_k$ follows from the fact that $\mathcal{F}_k$ is a ball with radius 2ε). Now, construct four of such collections of balls, $\{\mathcal{F}_k^q\}_{k=1}^{\nu_q}$ $(q = 1, \ldots, 4)$, corresponding to four probability measures $Q = \hat{F}_g \times \hat{F}_h$, $\hat{F}_g \times F_h$, $F_g \times \hat{F}_h$ and $F_g \times F_h$, respectively, and let $\tilde{\psi}_k^q$ be an element of $\mathcal{F}_k^q \cap \mathcal{F}^\lambda$ with $\int |\psi - \tilde{\psi}_k| dQ \leq 4\varepsilon$ for any $\psi \in \mathcal{F}_k^q$, where $\nu_q \leq A\varepsilon^{-V}$ for each q.

Given the four coverings $\{\mathcal{F}_k^q\}$ $(q = 1, \ldots, 4)$, we now show the uniform negligibility of $T_n^\lambda(\theta, \boldsymbol{\alpha})$. Look at the decomposition:

$$\begin{aligned}
&\sup_{(\theta,\boldsymbol{\alpha}) \in \Theta^\lambda \times \mathcal{A}} |T_n^\lambda(\theta, \boldsymbol{\alpha})| \\
= \ & \sqrt{n} \sup_{(\theta,\boldsymbol{\alpha}) \in \Theta^\lambda \times \mathcal{A}} \left| \int\int \psi^\lambda(x, \tilde{x}, \theta, \boldsymbol{\alpha}) [d\hat{F}_g d\hat{F}_h - d\hat{F}_g dF_h - dF_g d\hat{F}_h - dF_g dF_h] \right| \\
\leq \ & \sqrt{n} \Big[\max_{k \in \{1,\ldots,\nu_1\}} \sup_{\psi \in \mathcal{F}_k^1} \int\int |\psi - \tilde{\psi}_{1,k}| d\hat{F}_g d\hat{F}_h + \max_{k \in \{1,\ldots,\nu_2\}} \sup_{\psi \in \mathcal{F}_k^2} \int\int |\psi - \tilde{\psi}_{2,k}| d\hat{F}_g dF_h \\
& \max_{k \in \{1,\ldots,\nu_3\}} \sup_{\psi \in \mathcal{F}_k^3} \int\int |\psi - \tilde{\psi}_{3,k}| dF_g d\hat{F}_h + \max_{k \in \{1,\ldots,\nu_4\}} \sup_{\psi \in \mathcal{F}_k^4} \int\int |\psi - \tilde{\psi}_{4,k}| dF_g dF_h \Big] \\
& + \sqrt{n} \max_{q \in \{1,2,3,4\};\ k \in \{1,\ldots,\nu^q\}} \left| \int\int \tilde{\psi}_{q,k} [d\hat{F}_g d\hat{F}_h - d\hat{F}_g dF_h - dF_g d\hat{F}_h - dF_g dF_h] \right|.
\end{aligned} \tag{D.31}$$

The first term on the RHS is bounded by $\sqrt{n} \times 4 \times 4\varepsilon = 16\sqrt{n}\varepsilon$. Letting $\varepsilon = \sqrt{(\log n)/n}$,

$$\text{the first term on the RHS of (D.31)} = O(\sqrt{(\log n)/n}). \tag{D.32}$$

To consider a probability bound of the second term, we use the following exponential inequality: for any $a > 0$,

$$\Pr\left[\left|\int\int \tilde{\psi}_{q,k}[d\hat{F}_g d\hat{F}_h - d\hat{F}_g dF_h - dF_g d\hat{F}_h - dF_g dF_h]\right| > a\right] \leq 2\exp\{-a^2/[\bar{C}/n^2 + 8a/3]\}, \tag{D.33}$$

uniformly over q and k, for some constant $\bar{C}(> 0$; independent of q and $k)$, whose proof is provided below. Then, given $\varepsilon = \sqrt{(\log n)/n}$, we have

$$\begin{aligned}
&\Pr\left[\max_{q\in\{1,2,3,4\};\ k\in\{1,...,\nu^q\}}\left|\int\int \tilde{\psi}_{q,k}[d\hat{F}_g d\hat{F}_h - d\hat{F}_g dF_h - dF_g d\hat{F}_h - dF_g dF_h]\right| > a\right] \\
\leq\ & \sum\nolimits_{1\leq q\leq 4;\ 1\leq k\leq \nu^q} \Pr\left[\left|\int\int \tilde{\psi}_{q,k}[d\hat{F}_g d\hat{F}_h - d\hat{F}_g dF_h - dF_g d\hat{F}_h - dF_g dF_h]\right| > a\right] \\
\leq\ & 4A\varepsilon^{-V} \times 2\exp\left\{-\frac{a^2}{\bar{C}/n^2 + 8a/3}\right\} \leq 8A\left[(\log n)/n\right]^{-V/2}\exp\left\{-\frac{\bar{a}^2(\log n)}{\bar{C} + 8\bar{a}\times o(1)}\right\} \to 0,
\end{aligned}$$

where we have set $a = \bar{a}\sqrt{(\log n)/n^2}$ for a constant $\bar{a} > 0$ in the last line, and the convergence takes place for any $\bar{a}$ large enough as $n \to \infty$. This means that

$$\begin{aligned}
&\max_{q\in\{1,2,3,4\};\ k\in\{1,...,\nu^q\}}\left|\int\int \tilde{\psi}_{q,k}[d\hat{F}_g d\hat{F}_h - d\hat{F}_g dF_h - dF_g d\hat{F}_h - dF_g dF_h]\right| \\
=\ & O_P(\sqrt{(\log n)/n^2})
\end{aligned} \tag{D.34}$$

By (D.31), (D.32) and (D.34), we now obtain (D.25).

We next verify the stochastic equicontinuity of $\{J_n^\lambda(\theta, \boldsymbol{\alpha})\}$. Note that we can write

$$\sqrt{n}[M_{1,n}^\lambda(\theta, \boldsymbol{\alpha}) - \bar{M}^\lambda(\theta, \boldsymbol{\alpha})] = \sqrt{n}\int \psi_1^\lambda(x, \theta, \boldsymbol{\alpha}) d[\hat{F}_g - F_g],$$

and the uniform covering number of $\left\{\psi_1^\lambda(x, \theta, \boldsymbol{\alpha})\right\}_{(\theta,\boldsymbol{\alpha})\in\Theta^\lambda\times\mathcal{A}}$ satisfies

$$\sup_{Q_1} N(2\varepsilon, \{\psi_1^\lambda(x, \theta, \boldsymbol{\alpha})\}_{(\theta,\boldsymbol{\alpha})\in\Theta^\lambda\times\mathcal{A}}, L_1(Q_1)) \leq A\varepsilon^{-V}, \tag{D.35}$$

where the supremum is taken over any probability measures on $\mathbb{R}^{2d}$, and A, V are constants given in (D.29). This (D.35) can be checked by setting $dQ = dQ_1 \times dF_h$ in (D.29), since $\psi_1^\lambda(x, \theta, \boldsymbol{\alpha}) = \int \psi^\lambda(x, \tilde{x}, \theta, \boldsymbol{\alpha}) dF_h(\tilde{x})$. Given this bound (D.35), the I.I.D. condition (Assumption 2), and the uniform boundedness of ψ_1^λ, by Andrews (1994, Theorem 1), we can check that the stochastic process $\{\sqrt{n}\left[M_{1,n}^\lambda(\theta, \boldsymbol{\alpha}) - \bar{M}^\lambda(\theta, \boldsymbol{\alpha})\right]\}_{(\theta,\boldsymbol{\alpha})\in\Theta^\lambda\times\mathcal{A}}$ is stochastically equicontinuous (SE) with respect to the pseudo metric

$$\rho_1((\theta_1, \boldsymbol{\alpha}_1), (\theta_2, \boldsymbol{\alpha}_2)) := \{\int |\psi_1^\lambda(x, \theta, \boldsymbol{\alpha}) - \psi_1^\lambda(x, \theta, \boldsymbol{\alpha})|^2 dF_g(x)\}^{1/2}.$$

In the same way, we can also check that $\{\sqrt{n}\left[M_{2,n}^\lambda(\theta, \boldsymbol{\alpha}) - \bar{M}^\lambda(\theta, \boldsymbol{\alpha})\right]\}_{(\theta,\boldsymbol{\alpha})\in\Theta^\lambda\times\mathcal{A}}$ is also SE with respect to

$$\rho_2((\theta_1, \boldsymbol{\alpha}_1), (\theta_2, \boldsymbol{\alpha}_2)) := \{\int |\psi_2^\lambda(\tilde{x}, \theta, \boldsymbol{\alpha}) - \psi_2^\lambda(\tilde{x}, \theta, \boldsymbol{\alpha})|^2 dF_h(\tilde{x})\}^{1/2}.$$

From these, we can also see that $\sqrt{n}\{\left[M_n^\lambda(\theta,\boldsymbol{\alpha}) - M_{1,n}^\lambda(\theta,\boldsymbol{\alpha}) - M_{2,n}^\lambda(\theta,\boldsymbol{\alpha}) + \bar{M}^\lambda(\theta,\boldsymbol{\alpha})\right]\}_{(\theta,\boldsymbol{\alpha})\in\Theta^\lambda\times\mathcal{A}}$ is SE with respect to the pseudo metric $\rho = \rho_1 + \rho_2$, i.e., we now have obtained (D.26). It remains to show (D.33).

Proof of (D.33): Let

$$\begin{aligned}\eta^\lambda(i,j,\theta,\boldsymbol{\alpha}) &= \psi^\lambda(X_{g,i},X_{h,j},\theta,\boldsymbol{\alpha}) - \int\psi^\lambda(X_{g,i},\tilde{x},\theta,\boldsymbol{\alpha})\,dF_h(\tilde{x}) \\ &\quad -\int\psi^\lambda(x,X_{h,j},\theta,\boldsymbol{\alpha})\,dF_g(x) + \bar{M}(\theta,\boldsymbol{\alpha}),\end{aligned}$$

where we note that $\bar{M}(\theta,\boldsymbol{\alpha}) = \int\int\psi^\lambda(x,\tilde{x},\theta,\boldsymbol{\alpha})\,dF_g(x)\,dF_h(\tilde{x})$, $\left|\eta^\lambda(i,j,\theta,\boldsymbol{\alpha})\right| \le 8$ by the form of ψ^λ (given in (D.27)), and $\left\{\eta^\lambda(i,j,\theta,\boldsymbol{\alpha})\right\}$ independent over i and j. Then, we can write

$$\int\int\psi[d\hat{F}_g d\hat{F}_h - d\hat{F}_g dF_h - dF_g d\hat{F}_h - dF_g dF_h] = (1/n_g n_h)\textstyle\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\eta^\lambda(i,j,\theta,\boldsymbol{\alpha}), \quad \text{(D.36)}$$

for a generic element of $\psi\in\mathcal{F}^\lambda$, with some $(\theta,\boldsymbol{\alpha})$.

To obtain (D.33), we first compute the L_2-moment bound of (D.36). For notational simplicity, we here write $\eta(i,j) = \eta^\lambda(i,j,\theta,\boldsymbol{\alpha})$, suppressing the dependence on λ, θ and $\boldsymbol{\alpha}$. Then, we have

$$\begin{aligned}&E\left[\left|(1/n_g n_h)\textstyle\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\eta^\lambda(i,j,\theta,\boldsymbol{\alpha})\right|^2\right] \\ =\;& (1/n_g n_h)^2 E\left[\textstyle\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\sum_{k=1}^{n_g}\sum_{l=1}^{n_h}\eta(i,j)\eta(k,l)\right] \\ =\;& (1/n_g n_h)^2\Big\{\textstyle\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}E\left[|\eta(i,j)|^2\right] + \sum\sum\sum_{1\le i\le n_g;\ 1\le j,l\le n_h;\ j\ne l}E[\eta(i,j)\eta(i,l)] \\ &+\sum\sum\sum_{1\le i,k\le n_g;\ i\ne k;\ 1\le j\le n_h}E[\eta(i,j)\eta(k,j)] \\ &+\sum\sum\sum\sum_{1\le i,k\le n_g;\ i\ne k;\ 1\le j,l\le n_h;\ j\ne l}E[\eta(i,j)\eta(k,l)]\Big\} \\ =\;& (1/n_g n_h)^2\textstyle\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}E[|\eta(i,j)|^2], \qquad \text{(D.37)}\end{aligned}$$

where the last equality holds since

$$\begin{aligned}E[\eta(i,j)\eta(i,l)] &= 0 \ \text{ for } j\ne l; \quad E[\eta(i,j)\eta(k,j)] = 0 \ \text{ for } i\ne k; \qquad \text{(D.38)}\\ E[\eta(i,j)\eta(k,l)] &= 0 \ \text{ for } i\ne k \text{ and } j\ne l. \qquad \text{(D.39)}\end{aligned}$$

We can easily check (D.39) by the independence of $\eta(i,j)$ and $\eta(k,l)$: $E[\eta(i,j)\eta(k,l)] = E[\eta(i,j)]\,E[\eta(k,l)] = 0$. The two equalities in (D.38) follow from almost the same argument, and we here check only the former:

$$\begin{aligned}&E[\eta(i,j)\eta(i,l)] \\ =\;& E\left[\left\{[\psi^\lambda(X_{g,i},X_{h,j},\theta,\alpha) - \textstyle\int\psi^\lambda(X_{g,i},\tilde{x},\theta,\boldsymbol{\alpha})dF_h(\tilde{x})] - [\int\psi^\lambda(x,X_{h,j},\theta,\alpha)dF_g(x) - \bar{M}(\theta,\boldsymbol{\alpha})]\right\}\right. \\ &\left.\times\left\{[\psi^\lambda(X_{g,i},X_{h,l},\theta,\alpha) - \textstyle\int\psi^\lambda(X_{g,i},\tilde{x},\theta,\boldsymbol{\alpha})dF_h(\tilde{x})] - [\int\psi^\lambda(x,X_{h,l},\theta,\alpha)dF_g(x) - \bar{M}(\theta,\boldsymbol{\alpha})]\right\}\right] \\ =\;& 0,\end{aligned}$$

which holds by the mean-zero property of four components in parenthesis and the independence between $X_{g,i}$, $X_{h,j}$ and $X_{h,l}$. For each λ, $\eta(i,j) = \eta^{\lambda}(i,j,\theta,\boldsymbol{\alpha}) \leq 8$ uniformly over $(\theta,\boldsymbol{\alpha})$, and therefore $E[|\eta(i,j)|^2] = E[|\eta^{\lambda}(i,j,\theta,\boldsymbol{\alpha})|^2] \leq 64$ uniformly. Now, by (D.37) and Assumption 1, we have

$$\sup_{(\theta,\boldsymbol{\alpha})\in\mathbb{R}\times\mathbb{R}^{4d}} E\left[\left|(1/n_g n_h)\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\eta^{\lambda}(i,j,\theta,\boldsymbol{\alpha})\right|^2\right] \leq 64/n_g n_h \leq \bar{C}/n^2,$$

for some constant $\bar{C} > 0$.

Given this moment bound, we now apply the Bernstein inequality for independent and bounded sequences (e.g., Lemma 2.2.9 of van der Vaart and Wellner, 1996), and obtain

$$\Pr\left[\left|(1/n_g n_h)\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\eta^{\lambda}(i,j,\theta,\boldsymbol{\alpha})\right| > a\right] \leq 2\exp\{-a^2/[\bar{C}/n^2 + 8a/3]\},$$

where the inequality holds uniformly over $(\theta,\boldsymbol{\alpha}) \in \mathbb{R}\times\mathbb{R}^{4d}$ (since $\bar{C}$ is independent of $(\theta,\boldsymbol{\alpha})$), leading to the desired result (D.33). Now, the proof of Lemma 3 is completed.

References for Part D of Appendix

Andrews, D.W.K. (1994) Empirical process methods in econometric, Handbook of Econometrics, Vol IV., Chapter 37 (R. F. Engle & D. McFadden ed.), 2247-2294, Elsevier.

Chen, X., O. Linton & I. van Keilegom (2003) Estimation of semiparametric models when the criterion function is not smooth, Econometrica, 71-5, 1591-1608.

Kosorok, M.R. (2008) Introduction to Empirical Processes and Semiparametric Inference, Springer.

Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators, Econometrica, 57-5, 1027-1057.

Newey, W.K. (1984) A method of moments interpretation of sequential estimators, Economic Letters, 14, 201-206.

Newey, W.K. & D. McFadden (1994) Large sample estimation and hypothesis testing, Handbook of Econometrics, Vol IV (R. F. Engle & D. McFadden ed.), Chapter 36, 2111-2245, Elsevier.

Nolan, D. & D. Pollard (1987) U-processes: Rates of convergence, The Annals of Statistics, 15, 780-799.

Nolan, D. & D. Pollard (1988) Functional limit theorems for U-processes, The Annals of Probability, 16, 1291-1298.

Sherman, R.P. (1993) The limit distribution of the maximum rank correlation estimator, Econometrica, 61-1, 123-137.

Sherman, R.P. (1994) Maximal inequalities for degenerate U-processes with applications to optimization estimators, The Annals of Statistics, 22, 439-459.

Van der Vaart, A.W. & J. Wellner (1996) Weak Convergence and Empirical Processes: With Applications to Statistics, Springer.

Research Papers 2013

2013-41: Diego Amaya, Peter Christoffersen, Kris Jacobs and Aurelio Vasquez: Does Realized Skewness Predict the Cross-Section of Equity Returns?

2013-42: Torben G. Andersen and Oleg Bondarenko: Reflecting on the VPN Dispute

2013-43: Torben G. Andersen and Oleg Bondarenko: Assessing Measures of Order Flow Toxicity via Perfect Trade Classification

2013-44: Federico Carlini and Paolo Santucci de Magistris: On the identification of fractionally cointegrated VAR models with the F(d) condition

2013-45: Peter Christoffersen, Du Du and Redouane Elkamhi: Rare Disasters and Credit Market Puzzles

2013-46: Peter Christoffersen, Kris Jacobs, Xisong Jin and Hugues Langlois: Dynamic Diversification in Corporate Credit

2013-47: Peter Christoffersen, Mathieu Fournier and Kris Jacobs: The Factor Structure in Equity Options

2013-48: Peter Christoffersen, Ruslan Goyenko, Kris Jacobs and Mehdi Karoui: Illiquidity Premia in the Equity Options Market

2013-49: Peter Christoffersen, Vihang R. Errunza, Kris Jacobs and Xisong Jin: Correlation Dynamics and International Diversification Benefits

2013-50: Georgios Effraimidis and Christian M. Dahl: Nonparametric Estimation of Cumulative Incidence Functions for Competing Risks Data with Missing Cause of Failure

2013-51: Mehmet Caner and Anders Bredahl Kock: Oracle Inequalities for Convex Loss Functions with Non-Linear Targets

2013-52: Torben G. Andersen, Oleg Bondarenko, Viktor Todorov and George Tauchen: The Fine Structure of Equity-Index Option Dynamics

2014-01 Manuel Lukas and Eric Hillebrand: Bagging Weak Predictors

2014-02: Barbara Annicchiarico, Anna Rita Bennato and Emilio Zanetti Chini: 150 Years of Italian CO2 Emissions and Economic Growth

2014-03: Paul Catani, Timo Teräsvirta and Meiqun Yin: A Lagrange Multiplier Test for Testing the Adequacy of the Constant Conditional Correlation GARCH Model

2014-04: Timo Teräsvirta and Yukai Yang: Linearity and Misspecification Tests for Vector Smooth Transition Regression Models

2014-05: Kris Boudt, Sébastien Laurent, Asger Lunde and Rogier Quaedvlieg: Positive Semidefinite Integrated Covariance Estimation, Factorizations and Asynchronicity

2014-06: Debopam Bhattacharya, Shin Kanaya and Margaret Stevens: Are University Admissions Academically Fair?