Preliminary:
Comments welcome.

Detecting Discrimination in Audit and Correspondence Studies*

David Neumark

Department of Economics, UCI,
NBER, and IZA

June 28, 2010

Abstract

Audit studies to test for labor market discrimination have been criticized for failing to ensure that applicants from different groups appear identical to employers. Many of these criticisms can be countered by using correspondence studies with fictitious paper applicants whose qualifications on paper can be made identical across groups. However, Heckman and Siegelman (1993) show that even in the ideal correspondence study in which both observed and unobserved group averages are identical, group differences in the variances of unobservable determinants of productivity can lead to spurious evidence of discrimination or of reverse discrimination, or mask evidence of discrimination. This paper shows how an unbiased estimate of discrimination can be recovered even when there are group differences in the variances of the unobservables, as long as the correspondence study includes variation in observable measures of applicant quality that affect the probability of hiring. The method is applied to data from Bertrand and Mullainathan's (2004) correspondence study, and yields stronger evidence of race discrimination that adversely affects blacks than occurs when differences in the variances of the unobservables are ignored, hence strengthening their finding of race discrimination. The method proposed in this paper can be easily implemented in any future correspondence study that collects the requisite data.

* I am grateful to the UCI Academic Senate Council on Research, Computing, and Libraries for research support, to Marianne Bitler, Thomas Cornelissen, Judith Hellerstein, James Heckman, and seminar participants at Hebrew University and Tel Aviv University for helpful comments, and to Jennifer Graves, Smith Williams, and especially Andrew Chang for research assistance. I also thank Marianne Bertrand and Sendhil Mullainathan for supplying their data. An earlier version of this paper was presented at the 2010 SOLE/EALE meetings.

I. Introduction

In audit or correspondence studies, fictitious individuals who are identical except for race, sex, or ethnicity apply for jobs. Evidence of group differences in outcomes – for example, blacks getting fewer job offers than whites – is generally viewed as compelling evidence of discrimination. Across a wide array of countries and demographic groups, audit or correspondence studies conclude that there is evidence of discrimination, including, for example, discrimination against blacks, Hispanics, and women in the United States (Mincy, 1993; Neumark, 1996; Bertrand and Mullainathan [BM], 2004), Moroccans in Belgium and the Netherlands (Smeeters and Nayer, 1998; Bovenkerk et al., 1995), and lower castes in India (Banerjee et al., 2008). These "field experiments" for testing for discrimination are widely viewed as providing the most convincing evidence on discrimination (Pager, 2007; Riach and Rich, 2002), and U.S. courts allow organizations that conduct audit or correspondence studies to file claims of discrimination based on the evidence they collect (U.S. Equal Employment Opportunity Commission, 1996).[1]

Nonetheless, audit or correspondence studies have been subjected to scrutiny and criticism, most notably from Heckman and Siegelman [HS] (HS, 1993; Heckman, 1998). Most importantly, audit studies have been criticized for failing to ensure that applicants from different groups appear identical to employers. Many of these criticisms can be countered by using correspondence studies with fictitious paper applicants whose qualifications on paper can be made identical across groups. However, Heckman and Siegelman (1993) show that even in the ideal correspondence study in which both observed and unobserved group averages are identical, group differences in the variances of unobservable determinants

[1] EEOC (1996) discusses the case history regarding the use of testers in labor market and housing discrimination, as well as the Commission's conclusions. Interestingly, the EEOC refers to case law from the early civil rights movement when groups' only motivation was to test the law. For example, the EEOC cites *Pierson v. Ray*, 386 U.S. 547 (1967), in which the Supreme Court "held that a group of Black clergymen who were removed from a segregated bus terminal in Jackson, Mississippi had standing to seek redress," ruling that the plaintiffs "had been discriminated against by being ejected from the terminal, despite the fact that the plaintiffs' sole purpose was to test the law rather than to actually use the terminal" (U.S. EEOC, 1996, p. 2). The EEOC cites similar case law regarding plaintiffs testing the legality of segregation laws on buses. The document discusses the well-established standing of testers in housing discrimination cases under Title VIII, based on *Havens Realty Corp. v. Coleman,* 455, U.S. 363, 373 (1982), the parallels to employment discrimination under Title VII, and cases in which damages have been awarded in employment testing. Finally, it discusses *Fair Employment Council of Greater Washington, DC v. BMC Marketing Corporation*, 28 F 3d 1268 (D.C. Circuit 1994), in which standing was limited because individual testers could not show future harm, as well as the implications of Civil Rights Act of 1991, which allows for damages, according to the EEOC, would have given the testers in this case standing had the testing occurred after its passage (p. 4).

of productivity can lead to substantial spurious evidence of discrimination or of reverse discrimination. This critique is inherently difficult to address because it emphasizes factors outside the control of the researcher. As a consequence of this difficulty, apparently, this critique has been ignored in the literature, and researchers have continued to carry out additional field experiments of discrimination without regard to the strong possibility of highly misleading results. In fact, however, the HS critique poses a significant challenge to the audit/correspondence study method.

This paper explicitly addresses the HS "unobservable variance" critique. In particular, it develops and implements a general method of using data from correspondence studies that accomplishes two goals. First, it provides a statistical test of whether this HS critique applies to the data from any particular study. Second, and more important, if the HS critique does apply, it yields a statistical estimation procedure that uncovers an unbiased estimate of discrimination.[2] It is a simple matter to collect the requisite data in future correspondence studies, and the method is also implemented using data from a correspondence study (BM, 2004) that has the requisite data. Finally, the estimation procedure is assessed via Monte Carlo simulations.

II. Background on Audit and Correspondence Studies

Earlier research on labor market discrimination focused on individual-level employment or earnings regressions, with the estimate of discrimination inferred from the race, sex, or ethnic differential that remains unexplained after including a wide array of proxies for productivity. These analyses suffer from the obvious criticism that the proxies do not adequately capture group differences in productivity, in which case the "unexplained" differences cannot be interpreted as detecting discrimination.

Audit or correspondence studies are a response to this inherent weakness of the regression approach to discrimination.[3] These studies are based on comparisons of outcomes (usually job interviews or job offers) for matched job applicants of different races or sexes (see, e.g., Turner, et al., 1991; Neumark, 1996; BM, 2004). Audit or correspondence studies directly address the problem of missing data

---

[2] To clarify, it uncovers an unbiased estimate of discrimination in hiring (or callbacks for interviews) for the applicants in the study. There are other criticisms of audit and correspondence studies related to how generalizable the results are to the population, and what the results tell us about the existence of discrimination at the level of the market. See HS (1993) and Heckman (1998) for more details. This paper does not address those issues.
[3] Another approach is to try to incorporate data in which productivity is observed or can be estimated (e.g., Foster and Rosenzweig, 1993; Hellerstein et al., 1999).

on productivity. Rather than try to control for variables that might be associated with productivity differences between groups, these studies instead create an artificial pool of job applicants, among which there are intended to be no average differences by race, for example. By using either applicants coached to act alike, with identical-quality resumes (an audit study), or simply applicants on paper who have equal qualifications (a correspondence study), the method is largely immune to criticisms of failure to control for important differences between, for example, black and white job applicants. As a consequence, this strategy has come to be widely used in testing for discrimination in labor markets (as well as housing markets). Thorough reviews are contained in Fix and Struyk (1993) and Riach and Rich (2002).

Despite the widely-held view that audit or correspondence studies are the best way to test for labor market discrimination, some critiques of these studies challenge their conclusions (HS, 1993; Heckman, 1998). Some of these criticisms are acknowledged as potentially valid, and subsequent research has adapted. For example, HS noted that in the prominent audit studies carried out by the Urban Institute (e.g., Mincy, 1993), white and minority testers sent out to apply for jobs were aware of the purpose of the test, and even – in their training – informed about "the pervasive problem of discrimination in the United States," raising the possibility that testers subconsciously took actions in their job interviews that led to the "expected" result (HS, 1993). A constructive response to this criticism has been the move to correspondence studies, which focus on applications on paper and whether they result in job interviews, thus cutting out the influence of the individual job applicants used in the test.

However, a fundamental critique of audit or correspondence studies is one that has not been addressed by researchers. In particular, HS consider what most researchers view as the ideal conditions for an audit or correspondence study – when not only are the observable average differences between groups eliminated, but in addition the observable characteristics used in the applications are sufficiently rich that it is reasonable to assume that potential employers believe there no average differences in *unobservable* characteristics across groups. HS show that, even in this case, audit or correspondence studies can generate evidence of discrimination (in either direction) when there is none, and can also mask evidence of discrimination when it in fact exists. Given the pervasive use of audit and correspondence studies, the absence of any research assessing this critique is a significant gap in the social science and legal literature.

III. The Heckman-Siegelman Critique

Set-up with Continuous "Treatment" Decision

Suppose that productivity depends on two individual characteristics, $X' = (X^I, X^{II})$. Let R be a dummy for race, with R = equal to 1 for minorities and 0 for non-minorities (which I will refer to as "black" and "white" for short). Allow productivity also to depend on a firm-level characteristic F, so that productivity is P(X',F). Let the treatment of a worker depending on P and possibly R (if there is discrimination) be denoted T(P(X',F),R). For now, think of this treatment as continuous, even though that is not the usual outcome for an audit or correspondence study; suppose the treatment is, for example, the wage offered, set equal to a worker's productivity minus a possible discriminatory penalty for blacks as in Becker's (1971) employer taste discrimination model.

Define discrimination as

(1) $T(P(X',F)|R = 1) \neq T(P(X',F)|R = 0)$.

Assume that P(.,.) and T(P(.,.)) are additive, so

(2) $P(X',F) = X^I + X^{II} + F$

(3) $T(P(X',F),R) = P + \gamma R$.[4]

Thus, discrimination against blacks implies that $\gamma < 0$, so that blacks are paid less than equally-productive whites at the same firm.

In an audit or correspondence study, two testers or multiple pairs of testers (one with R = 1 and one with R = 0 in each pair) are sent to firms to apply for jobs. The researcher attempts to standardize their productivity based on observable productivity-related characteristics. Denote expected productivity for blacks and whites, based on what the firm observes, as P_B^* and P_W^*; note that I have not specified that these are necessarily based on both X^I and X^{II}, as we may want to treat X^{II} as unobserved by firms. The goal of the audit or correspondence study design is to set $P_B^* = P_W^*$. Given these observables, the outcome T is

[4] For now, I treat the productivity variables in the abstract, and therefore without loss of generality can assume them to be scaled such that their coefficients equal one. When I turn to the data and relate productivity-related characteristics to expected productivity, I will introduce coefficients multiplying the productivity-related characteristics.

observed for each tester. So based on equation (3), each test – thought of as the outcome of applications to a firm by one black and one white tester – yields an observation

(4) $T(P_B^*,1) - T(P_W^*,0) = P_B^* + \gamma - P_W^*$.

If $P_B^* = P_W^*$, then averaging across tests yields an estimate of γ. Alternatively, γ can be estimated from the regression

(5) $T(P_{ij}^*) = \alpha + \gamma R_i + \mu_j + \varepsilon_{ij}$

where $T(P_{ij}^*)$ is the outcome for worker i at firm j, R_i is a dummy variable for the race of worker i, and μ_j is a vector of firm fixed effects.[5]

Now consider explicitly the two observable components of productivity, X^I and X^{II}. Suppose the audit study controls only one of these; so think of X^I as the schooling level, which is controlled in the resumes/interviews, and X^{II} as another characteristic not controlled in the study and unobserved by employers. Denote by X_B^j and X_W^j the values of X^I and X^{II} for blacks and whites, j = I, II. Suppose that the audit or correspondence study, as is usually done, sets $X_B^I = X_W^I = X^{I*}$. Then for the test resulting from the application of a pair of black and white testers to a firm, P_B^* and P_W^*, the firm's expected productivity for workers in each group, are

(6) $P_B^* = X_B^I + E(X_B^{II}) + F$

(7) $P_W^* = X_W^I + E(X_W^{II}) + F$.

In this case, each individual test provides an observation equal to

(8) $T(P_B^*,1) - T(P_W^*,0) = P_B^* + \gamma - P_W^* = X_B^I + E(X_B^{II}) + \gamma - (X_W^I + E(X_W^{II}))$

$$= \gamma + E(X_B^{II}) - E(X_W^{II}).^6$$

[5] The inclusion of the firm fixed effects should have no implications for the estimate of γ, since the firm fixed effects are uncorrelated with race. Furthermore, when we introduce productivity-related characteristics X below, as long as testers' characteristics are randomly assigned to employers, which is typically the case in audit and correspondence studies, the estimated differentials associated with X do not depend on the inclusion of firm fixed effects. Thus, they are ignored in what follows. (Their inclusion does, however, affect the standard errors of the estimates, as the firm fixed effects may explain substantial amounts of variation in the outcome, and may also capture non-independence of the outcomes across testers at the same firm.)

[6] In the HS discussion, X^{II} is treated as uncontrolled in the research design but observed by employers. HS are in part concerned with observable differences between audit study testers that are not controlled, which makes sense in an audit study (although surely there are important unobservables that go undetected in the audit study job interviews). In a correspondence study, however, the employer observes *only* X^I, but may plausibly be assumed to know the

Mean Differences

To set the stage, I first discuss issues regarding mean differences across groups in observed and unobserved variables. Clearly observations from a sample of the tests described above provide an unbiased estimate of γ only if $E(X_B^{II}) = E(X_W^{II})$. Thus, a key assumption in an audit or correspondence study is that all productivity-related factors not controlled for in the test have the same mean for blacks and whites. Heckman (1998) and HS (1993, p. 245) point out, however, that researchers have limited information about what determines productivity within firms.[7] When there are uncontrolled productivity-related differences between black and white testers, not only will the audit or correspondence study produce a biased estimate of discrimination, but it can produce a *more* biased estimate than using randomly-matched pairs of testers. Suppose that, in addition to the preceding assumptions, X^I and X^{II} are statistically independent, and further assume that mean productivity is the same for blacks and whites,

(9) $E(P_B) = E(X_B^I + X_B^{II}) = E(P_W) = E(X_W^I + X_W^{II})$,

but that the mean of each characteristic differs by race, e.g., $E(X_W^I) \neq E(X_B^I)$. Coupled with equation (9), these assumptions imply that whites are more productive on one characteristic, and blacks on the other.

If an audit study controls for X^I but not X^{II}, the audit study estimates

(10) $\gamma + E(X_B^{II} - X_W^{II})$,

The bias here can be upward or downward, depending on whether workers are matched on characteristics on which blacks are more or less productive than whites. If the characteristic on which they are matched, X^I, is on average higher for blacks than for whites, then $E(X_B^{II} - X_W^{II}) < 0$ and the audit study

---

distribution of X^{II}. A natural question, then, is what generates variation in hiring across employers, if all employers simply evaluate applicants using the same means and variances of X_B^{II} and X_W^{II} for blacks and whites, respectively. Firm-specific differences in the productivity of any given worker can generate variation, so one way to think about race differences in the distribution of the unobservable is that the firm-specific productivity variable (F) has a component that has a different distribution for blacks and whites.

[7] Pager (2007) argues against this view, suggesting that "It is a mistake … to assume that the researcher is at a necessary disadvantage relative to the employer in identifying productivity-related characteristics" because "the researcher is herself or himself an employer in the planning and implementation of an audit study" (p. 115). That strikes me as an ineffective counter to the Heckman and Siegelman argument, as there is no reason to believe that what affects productivity in the firms in the audit study sample corresponds to what is important in hiring testers. Moreover, criteria other than productivity – in particular, the quality of matches between, for example, black and white testers – seem paramount in audit studies.

overstates discrimination, and conversely (discrimination against blacks implies $\gamma < 0$).[8] In this case, given the assumption that the expected *sum* of the productivity components is equal for blacks and whites (equation (9)), if we sent *randomly-matched* pairs of testers we would get an unbiased estimate of discrimination, because such an audit study would yield estimates of

(11) $$\gamma + E(X_B^I - X_W^I) + E(X_B^{II} - X_W^{II}) = \gamma + E(X_B^I + X_B^{II}) - E(X_W^I + X_W^{II}) = \gamma.$$ [9]

The previous case might be viewed as unrealistic, since its starting point is that blacks and whites are equally productive. Certainly much of the standard discrimination literature is premised on the possibility, at least, that blacks are less productive. If one adopts this view, then under the other assumptions outlined above, because the two expected differences in the first expression in equation (11) are no longer equal to zero, the audit study leads to less bias than using randomly-matched pairs. In this case, controlling for a larger number of productivity-related characteristics in an audit study can reduce the difference in uncontrolled productivity, and one might even believe that the remaining difference can be made sufficiently small that the bias in equation (11) becomes negligible.

However, that reasoning can break down if the assumption that X^I and X^{II} are statistically independent is dropped. Suppose, as before, that X^I and X^{II} each have different means for blacks and whites (and still that $E(X_B^{II} - X_W^{II}) < 0$), but now suppose they are *not* independent. The problematic case is when X^I and X^{II} are negatively correlated, in which case

(12) $$E(X_B^{II} - X_W^{II}) < E(X_B^{II} - X_W^{II} | X_B^I = X_W^I),$$

so that standardizing on X^I can accentuate the bias in an audit study, possibly leading to more bias from controlling for X^I rather than using randomly-matched testers – and possibly generating evidence of

[8] The converse case is emphasized by Darity and Mason (1998), who suggest that while whites have higher values of some of the usual productivity controls included in statistical studies of discrimination (such as schooling and AFQT), blacks have higher values of psychological variables such as self-esteem or locus of control, and of effort.

[9] This helps to demonstrate one of the main points of the Heckman/Siegelman critique with regard to mean differences between testers – that "Nowhere in the published literature on the audit pair method will you find a demonstration that matching one subset of observable variables necessarily implies that the resulting difference in audit-adjusted treatment between blacks and whites is an unbiased measure of discrimination – or indeed, that it is even necessarily a better measure of discrimination than comparing random pairs of whites and blacks applying at the same firm …" (Heckman, 1998, p. 108).

discrimination when γ is in fact zero.[10] This depends on the relative magnitudes of $\{E(X_B^I - X_W^I) + E(X_B^{II} - X_W^{II})\}$ – the bias with randomly-matched testers – and $E(X_B^{II} - X_W^{II} | X_B^I = X_W^I)$ – the bias with testers matched on X^I. The case where X^I and X^{II} are positively correlated may be more realistic, in which case standardizing on X^I implies that the difference in X^{II} between the black and white tester is on average *smaller* than for a randomly-selected pair.[11] Regardless, the test is still biased.

The discussion to here focuses on the problems that arise because audit or correspondence studies may not control for all characteristics that might differ by race. Good audit studies attempt to meet this criterion, but there are limits to how successfully they can do so.[12] Correspondence studies are a response to this potential problem, because, in contrast to audit studies, they do not entail face-to-face interviews that might convey mean differences on uncontrolled variables between blacks and whites.[13] This argument has become a common rationale for preferring correspondence studies to audit studies.

Even in a correspondence study, though, differences in employer estimates of mean unobserved characteristics for blacks and whites can affect the results, as in equation (10). The difference, in this case, is in part one of legal interpretation. In particular, because employers are not allowed to make assumptions about race (sex, etc.) differences in characteristics unobserved in the job application or interview process,[14] any role of assumed mean differences in characteristics in affecting the outcomes from a correspondence study can be interpreted as statistical discrimination. Consequently, we can interpret the estimate of the

---

[10] Moreover, in this case the role of relatively unimportant factors (yet still related to productivity) can be accentuated. Heckman and Siegelman (1993) discuss some examples in the context of the Urban Institute audit studies using Hispanic-white pairs.

[11] The argument is analogous to omitted variable bias in OLS regression. Suppose we are estimating the relationship between being black (B) and some outcome Y, and there are two omitted variables W and Z; ignore other covariates. Suppose W and Z are positively correlated with Y conditional on B (parallel to the example in the text, where the variables of interest, X^I and X^{II}, are positively related to productivity), but they have opposite-signed correlations with B. Then omitting both can lead to a less-biased estimate of the coefficient of B than including one but not the other. However, if all else is the same but W and Z have the same-signed correlation with B, then including one of them in the regression will unambiguously reduce the bias.

[12] As such, the statement in Hellerstein and Neumark (2006), that "The audit study approach … creates an artificial pool of labor market participants among whom there are no average differences by race or sex …" (p. 34) should be tempered. An audit study *attempts* to do this.

[13] This is a common argument in favor of correspondence studies, although they pose other disadvantages (e.g., Bertrand and Mullainathan, 2004, p. 994; Riach and Rick, 2002, p. F485). In addition, an audit study can explicitly compare results based on the application and interview stage, to see whether there is less discrimination at the application stage when face-to-face contact has been avoided (e.g., Neumark, 1996).

[14] See http://www.eeoc.gov/facts/fs-race.pdf (viewed March 23, 2009).

expression in equation (10) from a correspondence study as capturing the combined effects of taste discrimination (γ) and statistical discrimination ($E(X_B^{II} - X_W^{II})$).[15] That is, on legal grounds one might argue that correspondence studies, as opposed to audit studies, provide unbiased estimates of discrimination. According to this view, correspondence studies still do not, however, necessarily do any better at isolating taste discrimination – i.e., the discrimination that would remain if the means of all productivity-related characteristics were the same for blacks and whites. Moreover, economists are inherently interested in whether employers discriminate against groups with equal observed and equal expected unobserved characteristics, and the preceding discussion implies that correspondence studies may not provide an unbiased estimate of this more fundamental type of discrimination.

Distributional Differences in the Context of Hiring

A more troubling result emerges once we take account of the fact that, in audit or correspondence studies, the relevant treatment is not linear in productivity as it might be for a wage offer – like in equation (3) – but instead is non-linear. That is, we think that in the hiring process firms evaluate a job applicant's productivity relative to a standard, and offer the applicant a job (or an interview) if the standard is met. In this case, HS show that, even when the identifying assumption of equal group averages of *all* variables (observed or unobserved) holds, an audit or correspondence study can generate biased estimates, with spurious evidence of discrimination in either direction, or of its absence. Because this critique applies even to correspondence studies, which meet higher standards of validity – and applies even in the ideal case where there is *no* difference in the means of unobserved productivity measures – the remainder of the discussion refers exclusively to correspondence studies.

The intuitive basis of the HS critique is as follows. Consider the simplest case in which the only difference between blacks and whites is that the *variance* of unobserved productivity is higher for whites than for blacks, for example. That is, we assume $E(X_B^j) = (X_W^j)$, j = I, II. Imagine a correspondence study that controls for one productivity-related characteristic, X^I, and standardizes on a quite low value of X^I (that

---

[15] This is noted in the literature. See, for example, Heckman and Siegelman (1993), Riach and Rich (2002), Bertrand and Mullainathan (2004), and Lahey (2008). The latter three studies discuss the use of other information to try to distinguish between the two hypotheses.

is, the study makes the two groups equal on characteristic X^I, but at a low value X^{I*}). The correspondence study does not convey any information on a second, unobservable productivity-related characteristic, X^{II}. Because employers will offer a job interview only if the expected *sum* of $X^I + X^{II}$ is high, if X^{I*} is set at a low level, the employer has to believe that X^{II} is quite high in order to offer an interview. Even though the employer does not observe X^{II}, if the employer knows that the variance of X^{II} is higher for whites, then the employer correctly concludes that whites are more likely than blacks to have a sufficiently high sum of $X^I + X^{II}$, by virtue of the simple fact that fewer blacks have very high values of X^{II}. Employers will therefore be less likely to offer jobs to blacks than to whites, *even though the observed average of X^I is the same for blacks and whites, as is the unobserved average of X^{II}*. (The opposite holds if the standardization is at a high value of X^I; in the latter case the employer only needs to *avoid* very low values of X^{II}, which will be more common for whites.)

It is worth pointing out that the idea that the variances of unobservables differ across groups has a long tradition in research on discrimination, stemming from early models of statistical discrimination. For example, Aigner and Cain (1977) discuss these models and suggest that a higher variance of unobservables for blacks compared to whites is plausible, and Lundberg and Startz (1983) study how such an assumption can lead to an equilibrium with lower investment in human capital by blacks. On the other hand, Neumark (1999) finds no evidence that employers have better labor market information about whites than blacks; if anything, the point estimates go in the opposite direction, although the estimates are imprecise.

To see the bias result formally in our simple framework, suppose that a job offer or interview is given if a worker's expected productivity exceeds a certain threshold c. As before, suppose that P is determined as a linear sum of X^I, X^{II}, and F (equation (2)), with X^{II} (and F) statistically independent of X^I,[16] and the correspondence study controls for X^I. The hiring rules for blacks and whites (with the possibility of discrimination) are

(13) $T(P(X',F)|R = 1) = 1$ if $\beta_I X_B{}^I + X_B{}^{II} + \gamma + F > c$

[16] Again, we can treat F as statistically independent because resume characteristics are assigned randomly. X^{II} may not in fact be independent of X^I, but we can always think about X^{II} as the variation in the uncontrolled characteristic that is orthogonal to X^I.

(13') $T(P(X',F)|R = 0) = 1$ if $\beta_I X_W^I + X_W^{II} + F > c$.

Note that now X^I is now multiplied by a coefficient β_I, because in this section I discuss estimating models corresponding to equations (13) and (13') using data on an observable productivity variable X^I. Given that X^{II} is unobserved, its coefficient can be standardized to equal one.[17] Discrimination leads employers to "discount" the productivity of a black worker, as captured in γ.

Assume that the audit study controls for X^I, with $X_B^I = X_W^I = X^{I*}$. Assume further that X_B^{II} and X_W^{II} are normally distributed, with equal means (set to zero, without loss of generality), and standard deviations σ_B^{II} and σ_W^{II}. Finally, as long as the firm-specific productivity shifters F are normally distributed and independent of X^{II}, and have the same distribution for blacks and whites, then we can ignore them and focus solely on the variation in X^{II}.[18] Under these assumptions, the probabilities that the left-hand expressions in (13) and (13') equal one (blacks and whites get hired) are

(14) $\Pr[T(P(X^{I*},X_B^{II})|R = 1) = 1] = 1 - \Phi[(c - \beta_I X^{I*} - \gamma)/\sigma_B^{II}] = \Phi[(\beta_I X^{I*} + \gamma - c)/\sigma_B^{II}]$

(14') $\Pr[T(P(X^{I*},X_W^{II})\,|R = 0) = 1] = 1 - \Phi[(c - \beta_I X^{I*})/\sigma_W^{II}] = \Phi[(\beta_I X^{I*} - c)/\sigma_W^{II}]$,

where Φ denotes the standard normal distribution function.

The key to a correspondence study is that the difference between the two expressions in equations (14) and (14') – the success rates for black and white job applicants – is informative about discrimination. However, even if $\gamma = 0$, so there is no discrimination, these two expressions need not be equal because σ_B^{II} and σ_W^{II}, the standard deviations of X_B^{II} and X_W^{II}, can be unequal. It is possible to say something more precise. In particular, consider the earlier case with $\gamma = 0$, but $\sigma_W^{II} > \sigma_B^{II}$ – that is, the "uncontrolled" productivity-related variable has a larger standard deviation for whites than for blacks – and X^{I*} is set at a low level – i.e., the "standardization level" is low. Then the study will generate spurious evidence of discrimination against blacks. In particular, when $\beta_I X^{I*} < c$, $\sigma_W^{II} > \sigma_B^{II}$ and $\gamma = 0$ imply that the probability

---

[17] Following HS, I assume that the coefficient on X_1 is the same for blacks and whites, so the discrimination is reflected only in an intercept difference. I return to this issue later.

[18] That is, one can redefine the random variable in what follows as $X^{II} + F$, and the same reasoning goes through.

that blacks are hired is lower than the probability that whites are hired, and conversely when $\beta_I X^{I*} > c$.[19]

Thus, even if the means of the unobserved productivity-related variables are the same for each group, and firms use the same hiring standard for each group (i.e., $\gamma = 0$), audit studies can generate evidence consistent with discrimination against blacks (or, alternatively, in their favor).[20] Different combinations of the relative magnitudes of σ_W^{II} and σ_B^{II}, and whether the standardization value of X^{I*} is high or low, can generate similar or opposite results. This is the basis for HS's claim that even under ideal conditions correspondence (or audit) studies are uninformative about discrimination.

IV. Detecting Discrimination

With the right data from a correspondence study, and using the framework from the preceding section, we can test explicitly whether a correspondence study is biased in favor or against finding evidence of discrimination, and – going further – recover an unbiased estimate of discrimination (γ).

The intuition is as follows. The HS critique rests on differences between blacks and whites in the variances of unobserved productivity. The fundamental problem, as equations (14) and (14') illustrate, is that we cannot separately identify γ and a difference in the variance of the unobservables ($\sigma_B^{II}/\sigma_W^{II}$). But a higher variance for one group (say, whites) implies a smaller effect of observed characteristics on employment for that group. To see this in the limit, suppose that the variance of unobserved X^{II} were infinite for whites. Then X^I, the observed productivity-related variable, would have no effect on whether or not an employer thinks an applicant from that group meets the standard for hiring, given by equations (13) and (13'). Thus, information from a correspondence study on how variation in observable qualifications is related to employment outcomes can be informative about the relative variances of the unobservables, and this, in turn, can identify γ.

---

[19] Heckman (1998, footnote 7) suggests that the case with a low level of standardization and higher dispersion for whites "seems to rationalize" audit study evidence consistent with discrimination against blacks. It is not clear, however, that we know much about either the level of standardization or the relative dispersion of unobserved productivity. Even though the first issue relates to observables, there is no obvious way to compare the distributions of qualifications of testers in an audit study and of the relevant population of job applicants. In fact, there are two conflicting tendencies in setting standards for audit studies. Setting a low standard implies that call-back rates will be low, reducing the statistical power of the evidence. But setting a standard too high raises concerns about "overqualification" of candidates (e.g., BM, 2004, p. 995), which in an economic context presumably means that the employee will get a better job offer and hence will not take a job at the employer included in the test.

[20] This argument does not depend on normality. It will hold for symmetric distributions (Heckman, 1998).

More formally, consider a correspondence study with the assumptions from the previous section holding. Equations (14) and (14') imply that the difference in outcomes between blacks and whites is

(15) $\Phi[(\beta_I X^{I*} + \gamma - c)/\sigma_B^{II}] - \Phi[(\beta_I X^{I*} - c)/\sigma_W^{II}]$.

As equation (15) shows, the data cannot separately identify σ_B^{II} and σ_W^{II}, but rather only β_I/σ_B^{II} and β_I/σ_W^{II}. But that is not surprising. The two expressions in equation (15) are hiring probits for blacks and whites, respectively, with intercepts $(\gamma - c)$ for blacks and $-c$ for whites. In probit models (or similar latent variable models like logit) the coefficients are identified only up to a scale factor (which is why, typically, the variance is normalized to one in a probit model). To simplify, we can normalize one of the two variances above to equal one. Assume we do this for σ_W^{II}, and define σ_{BR}^{II} as the *relative* variance for blacks ($\sigma_{BR}^{II} = \sigma_B^{II}/\sigma_W^{II}$), so that equation (15) becomes

(15') $\Phi[(\beta_I X^{I*} + \gamma - c)/\sigma_{BR}^{II}] - \Phi[\beta_I X^{I*} - c]$.

This changes nothing, as γ is still unidentified. We cannot tell whether the intercept in the first cumulative normal in equation (15) differs because $\gamma \neq 0$ or because σ_{BR}^{II}, the relative variance for blacks, is different from 1. However, if there is variation in the level of qualifications used as controls (X^{I*}), and these qualifications affect hiring outcomes, then we can identify β_I/σ_{BR}^{II} and β_I in equation (15'), and the ratio of these two estimates provides an estimate of σ_{BR}^{II}.[21] And if we do inference on this ratio, we can test the hypothesis of equal standard deviations (or variances) of the unobservables. Finally, identification of σ_{BR}^{II} implies identification of γ. Note that without meaningful variation in X^{I*} this is not possible, since in that case all we have in the model are different intercepts with different parameters in both the numerators and the denominators ($(\gamma - c)/\sigma_{BR}^{II}$ and c).

Of course a critical assumption is that β_I is the same for blacks and whites. Otherwise, the ratio of the two coefficients does not identify σ_{BR}^{II}. As HS point out, the constancy of β_I is assumed in the Urban Institute studies that they critique, with discrimination entering through an intercept shift in the evaluation

---

[21] The discussion here is in terms of probit estimates of callbacks. It could just as well be couched in terms of logit estimation. Although typically (e.g., Maddala, 1983) the logit model is not written with the standard deviation of the error term appearing, it is possible to rewrite it in this way, in which case the difference in coefficients would again be informative about the ratio of the variances of the unobservable (Johnson and Kotz, 1970, p. 5).

of a worker's productivity, depending on their race. In the real world it is not hard to come up with reasons why the coefficients relating X^I to productivity might differ by race. For example, blacks and whites on average attend different schools, and if white schools are higher quality a given number of years of schooling may do more to increase white productivity than black productivity. But in a correspondence or audit study, it might be possible to control for these kinds of differences; for example, in this case one can control the area where applicants live, and hence hold school district constant (e.g., BM, 2004).

HS raise other possibilities. One is that there may be discrimination in evaluating particular attributes of a group. For example, employers may discriminate against high-education blacks but not low-education blacks. It is not possible to rule out differences in coefficients arising for these reasons. Finally, HS also suggest that differences in coefficients may reflect "statistical information processing" in the event of incomplete information about productivity, as in statistical discrimination models. Of course this is the idea underlying the identification strategy suggested above, as the difference in β_1 for blacks and whites is assumed to reflect precisely the accuracy with which X_1 signals productivity for each race. However, as discussed below, when there is data on multiple productivity-related characteristics there is more one can do to test whether there is homogeneity in the coefficients that makes it possible to identify $\sigma_{BR}{}^{II}$.

The estimation of $\beta_I/\sigma_{BR}{}^{II}$ and β_I, and inference on their ratio ($\sigma_{BR}{}^{II} = \sigma_B{}^{II}/\sigma_W{}^{II}$), can be done via a heteroskedastic probit model (e.g., Williams, 2009), which allows the variance of the unobservable to vary with race. To do this, we pool the data for blacks and whites. Similar to equation (5), there is a latent variable for perceived productivity assumed to be generated by:

(16) $T(P_{ij}{}^{*}) = -c + \beta_I X_{ij}{}^{I*} + \gamma R_i + \varepsilon_{ij}.$

Consistent with the earlier discussion, the error term ε captures the uncertainty from two sources: the firm fixed effect F assumed to be statistically independent of X^I and X^{II}, and, most important, the unobservable X^{II}. As is standard it is assumed that $E(\varepsilon_{ij}) = 0$, but the variance is assumed to follow:

(17) $\mathrm{Var}(\varepsilon_{ij}) = [\exp(\mu + \omega R_i)]^2.$

This model can be estimated via maximum likelihood. The observations should be treated as clustered on firms to obtain a variance-covariance matrix that is robust to the dependence of observations across firms. We normalize $\mu = 0$, given that there is an arbitrary normalization of the scale of the variance of one group (in this case whites, with $R_i = 0$). And the estimate of $\exp(\omega)$ is exactly the estimate of σ_{BR}^{II}.

We observe the application resulting in a hire if the expression in equation (16) exceeds zero, and otherwise not. In this model, what coefficients are identified? As it turns out, once we estimate this model this way (maintaining the assumption that β_I is the same for blacks and whites), γ is identified. To see this, observations on whites identify $-c$ and β_I, and observations on blacks identify $(-c + \gamma)/\exp(\omega)$ and $\beta_I/\exp(\omega)$. Thus, the ratio of $\beta_I/\{\beta_I/\exp(\omega)$ identifies $\exp(\omega)$, which, from equation (17), is the ratio of the standard deviation of the unobservable for blacks relative to whites and is, as before, identified from the ratio of the effect of X^{I*} on blacks relative to its effect on whites.[22] Once we have an estimate of $\exp(\omega)$ (or equivalently σ_{BR}^{II}), then, along with the estimate of c identified from whites, the expression $(-c + \gamma)/\exp(\omega)$ identified from blacks allows us to identify γ as well.[23]

Note that evidence on the σ_{BR}^{II} is itself informative. If it equals one, then there is no bias from differences in the distribution of unobservables. This would imply that a correspondence study, at least, is free from bias (under the other maintained assumptions). Alternatively, if σ_{BR}^{II} is not equal to one, but if we had some evidence on how the level of standardization X^{I*} compares to the relevant population of job applicants, we could determine the direction of bias. In particular, if the study *detects* discrimination and there is a bias *against* this finding – based on the estimate of the ratio of variances and information about X^{I*}, then the evidence of discrimination is not spurious, because it would be even stronger absent this bias.

---

[22] The fact that the estimate of σ_{BR}^{II} is the inverse of the estimate of the effects of X^{I*} on the probability of being hired (β_I/σ_B^{II} or β_I) makes sense. For example, when the variance of the unobservable is larger for blacks, a given change in X^{I*} has a smaller effect on the probability that a black is hired than on the probability that a white is hired, because this change in X^{I*} is less informative about black productivity, which is the sum $X^I + X^{II}$. This also clarifies why the availability of data with multiple values of X^I is essential to estimating the relative variance of the unobservables.

[23] Consistent with the earlier discussion of statistical discrimination, we might want to allow for the possibility that $E(X_B^{II} - X_W^{II}) \neq 0$, that is, that there is a difference in the mean of the unobservable between blacks and whites. In this case, we can normalize by assuming $E(X_W^{II}) = 0$ and defining $E(X_B^{II} - X_W^{II}) = \mu_{BW}^{II}$. In this case, we can replace γ in the preceding identification argument with $\gamma + \mu_{BW}^{II}$, and it is this sum of parameters, reflecting the combination of taste discrimination and the expected mean differences in the unobservable, that is identified. Thus, taste discrimination is identified only if we assume no mean difference in unobservables – an assumption that is untestable but can perhaps be made more plausible by including a rich set of background variables in the resumes used in the correspondence study. (This is more problematic in correspondence studies of age discrimination, because when a researcher gives older applicants the same amount of experience as younger applicants, there is a good chance that employers will make adverse assumptions about older applicants whose resumes reflect limited work experience.)

But because we can identify γ directly under the assumptions above, we can recover an estimate of discrimination that is not biased by the difference in the variances of the unobservables. And we can do this *without* having to make a determination as to whether X^{I*} used in the study is a high or low standardization; this is important because it may be impossible to establish the latter.

Thus, given the right data, we *can* determine whether a race difference in outcomes in a correspondence study in fact reflects discrimination, under the assumptions for which HS claimed this could not be done. No doubt those assumptions are restrictive, and there are almost surely ways to relax these assumptions and render the data uninformative about γ. But the nature of this approach is structural, and therefore of necessity rests on restrictions regarding parameters and functional forms.

Finally, we noted above that this method only works if β_I is the same for blacks and whites, since otherwise we cannot identify $\sigma_{BR}{}^{II}$ and hence cannot identify γ. If there is only one productivity control variable in the model, then all one can resort to is an assumption. However, with more controls, there is a testable restriction.[24] In particular, if there are multiple productivity controls, then if their coefficients differ between blacks and whites *only* because of the difference in the variance of the unobservable, the ratio of the coefficients between blacks and whites, for each variable, should be the same.[25] Consider that case with two observables X^I and Z^I. We modify equation (15') to be

(15'') $\Phi[(\beta_I{}^B X^{I*} + \beta_{II}{}^B Z^{I*} + \gamma - c)/\sigma_{BR}{}^{II}] - \Phi[(\beta_I{}^W X^{I*} + \beta_{II}{}^W Z^{I*} - c)]$.

The coefficients on the observables now have B and W superscripts to denote that they can differ by race. But if the only reason the coefficients differ from by race is because $\sigma_{BR}{}^{II} \neq 1$, then it must be the case that

(18) $\beta_I{}^B/\beta_I{}^W = \beta_{II}{}^B/\beta_{II}{}^W$.

Thus, we can test whether the restriction that is implied by homogeneity of effects but unequal variances of the unobservables holds. Of course we cannot literally rule out the possibility that the

[24] And, as noted earlier, with richer controls homogeneity of the coefficients might be more likely anyway.

[25] The possibility of differences in coefficients owing to statistical discrimination could also generate differences in coefficients between blacks and whites, and this could vary for different characteristics to the extent that the signal content of these characteristics varies. The argument below applies to this case as well.

coefficients differ by race, $\sigma_{BR}^{II} = 1$, *and* equation (18) holds, in which case there is no "single" estimate of the discriminatory differential in hiring, but discrimination may vary with the values of X and Z.[26] However, as we increase the number of control variables, it becomes increasingly likely that a restriction like equation (18) would hold for the whole set of coefficients only because of a difference in the variance of the unobservables. It is also possible to choose – as an identifying assumption – a subset of the observable characteristics (e.g., only X^I in equation (15')) for which equation (18) holds, and to identify σ_{BR}^{II} only from the coefficients of this subset of variables, or to test the restriction on the other coefficients as overidentifying restrictions.

A final issue concerns the interpretation of the coefficients from the probit or heteroskedastic probit models. As usual, to translate probit coefficient estimates into magnitudes that can be interpreted as the marginal effects of a variable (Z_k, generically, with coefficient β_k, when Z is the vector of controls with coefficients β), we use

(19) $\partial P(\text{hire})/\partial Z_k = \beta_k\phi(Z\beta)$,

where $\phi(.)$ is the standard normal density, and the standard deviation of the unobservable is normalized to one. Typically this is evaluated at the means of Z. When Z_k is a dummy variable – such as race, which is the focus of this paper – the difference in the cumulative normal distribution functions is often used instead, although the difference is usually trivial.

In the case of the heteroskedastic probit model, the issue is more subtle, because if the variances of the unobservables differ by race, then when race "changes" both the variance and the level of the latent variable (the valuation of a worker's productivity) that determines hires can shift. What we are interested in comparing with the expression in (19), however, is the latter effect only, since – and this is the whole point of the HS critique – differential treatment of blacks and whites based only on differences in variances of unobservables should not be interpreted as discrimination. Thus, we want to isolate the effect of race on the valuation of the worker. As long as we use the continuous version of the partial derivative (equation

---

[26] This is exactly the same issue that arises in estimating wage discrimination using a single dummy variable for race with the coefficients of the other variables the same for blacks and whites, as opposed to an Oaxaca-type (1973) decomposition.

(19) above), there is a natural decomposition of the effect of a change in Z_k into these two components.

When equation (17) is generalized to

(17') $\mathrm{Var}(\varepsilon_{ij}) = [\exp(W\omega)]^2$,[27]

where the vector of variables W includes Z, with coefficient ω_k, then the overall partial derivative of P(hire) with respect to Z_k is

(20) $\partial P(\text{hire})/\partial Z_k = \phi(Z\beta/\exp(W\omega))\cdot\{(\beta_k - Z\beta\cdot\omega_k)\}/\exp(W\omega)$.[28]

This expression can be broken into two pieces. First, the partial derivative with respect to changes in Z_k affecting only the level of the latent variable – corresponding to the counterfactual of Z_k changing the valuation of the worker without changing the variance of the unobservable – is equal to

(20') $\phi(Z\beta/\exp(W\omega))\cdot\{\beta_k/\exp(W\omega)$.

Second, the partial derivative with respect to changes via the variance of the unobservable is equal to

(20'') $\phi(Z\beta/\exp(W\omega))\cdot\{(-Z\beta\cdot\omega_k)\}/\exp(W\omega)$.

In the analysis below, these two separate effects are reported as well as the overall marginal effect; standard errors are calculated using the delta method.

V. Evidence, Implementation, and Assessment

Existing Evidence

As the preceding discussion shows, we need information on the effects of productivity-related characteristics on hiring or callbacks, estimated separately for blacks and whites (or other groups included in correspondence studies). Reporting of such results is rare in the literature. However, BM (2004) report race differences in the effects of different qualifications, although their interest has nothing to with the concerns of this paper, but rather is for purposes of inferring whether blacks have different incentives to

[27] Recall that μ in equation (17) is normalized to zero.
[28] See Cornelißen (2005).

invest in higher qualifications than do whites.[29]

Their conclusions are based on callback differences between the resumes they constructed to be low versus high quality. They find that white callback rates are higher for both types of resumes, and that white callback rates increase significantly with resume quality (from 8.5 to 10.8 percent). In contrast, black callback rates increase only slightly (from 6.2 to 6.7 percent), and the change is not statistically significant, leading them to conclude that, "African-Americans experience much less of an increase in callback rates for similar improvements in their credentials" (pp. 1000-1). Similar qualitative conclusions are reached based on an analysis that measures resume quality for one part of the sample by using an equation for the probability of callbacks estimated from another part of the sample. In this analysis, both groups experience an increase in callback rates from higher-quality resumes, but the effect is larger for whites.

Finally, and most pertinent to this paper, they report estimated probit models, for whites and blacks separately (their Table 5). These estimates reveal substantially stronger effects of measured qualifications on whites than on blacks. Among the estimated coefficients that are statistically significant for at least one group, this is true for experience,[30] having an email address, working while in school, academic honors, and other special skills (such as language). The only exception is for computer skills, which inexplicably have a negative effect on callback rates for whites.[31]

As the present paper suggests, an alternative interpretation of smaller estimated coefficients for blacks than for whites is a difference in the variance of the unobservables. In particular, the lower coefficients for blacks are consistent with a larger variance for blacks, i.e., $\sigma_{BR}^{II} > 1$. If it is also true that this correspondence study has chosen low levels of the control variables on which to standardize applicants, then the HS analysis would imply that there is a bias towards finding discrimination *in favor of* blacks. BM explicitly state that they tried to avoid overqualification even of the *higher-quality* resumes (p. 995), and if the qualifications of applicants were in fact low relative to the population of applicants, then BM's

[29] Their study actually concerns differences in treatment between black-sounding names and names that do not sound black. For simplicity, I discuss the results as if they capture differences between blacks and whites, which is certainly a plausible interpretation of their findings.
[30] This variable enters as a quadratic, and the effect of experience is stronger for whites up to about 16 years of experience. This is more than twice the mean in their sample.
[31] Also, the effect of gaps in employment is inexplicably positive, but not significant when we disaggregate by race.

evidence pointing to discrimination against blacks would be even stronger absent the bias from differences in the distribution of unobservables. Of course it is very difficult to know or assess whether the characteristics of applicants were low, since there is no way to identify the population of applicants. Hence, implementation of the estimation procedure proposed in this paper is likely the only way even to *sign* the bias, let alone to recover an unbiased estimate of discrimination. Nonetheless, if BM did use a low level of standardization, then the corrected estimate should reveal stronger evidence of discrimination against blacks.

Implementation Using Bertrand and Mullainathan Data

Using the methods described earlier with BM's data, it is possible both to test whether the differences in coefficient estimates are consistent with a simple difference in the variance of unobservables, and it is possible to uncover an unbiased estimate of discrimination (γ). This analysis is reported in Tables 1 and 2, using their data. Table 1 begins by simply presenting probit estimates for the probability of a callback. Marginal effects are reported,[32] and specifications are reported with no controls except a dummy variable for females (in columns (1)-(3)), adding controls for the individual characteristics included on the resumes, and finally adding also neighborhood characteristics for the applicant's zip code. The specific variables are listed in the footnote to the table. Estimates are shown for males and females combined, and for females only; as the sample sizes indicate, the male sample is considerably smaller.[33] Aside from the estimated effects of race, estimated effects are shown for a few of the resume characteristics capturing applicants' qualifications.

Echoing BM's conclusions, there is a sizable and statistically significant difference between the callback rates for blacks and whites, with the rate for blacks lower by 3-3.3 percentage points (or about 33 percent relative to the white callback rate of 9.65 percent). The estimated race differences are robust to the inclusion of the different sets of control variables, which is what we should expect in a correspondence

[32] The marginal effects are calculated by comparing the probability of a callback at the means of the variables, for each group separately, which corresponds to the difference in equation (15).

[33] These specifications estimated for males yielded similar results for the effects of race, although the estimated coefficients of some of the productivity-related characteristics were quite imprecise or had unexpected signs. In estimating the heteroskedastic probit model for males, in some cases there were computational problems, likely reflecting these other issues regarding the estimates for males, and perhaps also the smaller sample for males.

study in which the resumes are assigned randomly.

Interestingly, in light of the results of other audit and correspondence studies, there is no evidence of lower (or higher) callback rates for females than for males.[34] At the same time, the table also shows that a number of the resume characteristics have statistically significant effects on the callback probability; this, of course, is an essential input for using the methods described above to recover an unbiased estimate of discrimination.

The main analysis is reported beginning in Table 2. The analysis is presented for the specifications with the full set of individual resume controls, and then adding as well the full set of neighborhood controls. Panel A simply repeats the estimated race effects from Table 1, to make comparisons easier. Panel B begins by reporting the estimated marginal effects of race from the heteroskedastic probit model. These specifications recover the unbiased estimate of the effect of race discrimination, conditional on the restriction that the ratios of coefficients for blacks to those for whites are equal (equation (18)). As the table shows, these estimates are slightly smaller (in absolute value) than the estimates from the simple probits, but trivially so. They remain statistically significant, and indicate callback rates that are lower for blacks by about 2.4-2.5 percentage points (or about 25 percent).

However, these effects represent the effects on both the level of the latent variable (the valuation of the worker's productivity and the variance of the unobservable. Decomposing the effect of race, the effect via the level of the latent variable is larger than the marginal effect from the probit estimation, ranging from −0.054 to −0.086. The effect of race via the variance of the unobservable, in contrast, is positive, ranging from 0.028 to 0.062. (This latter effect is not statistically significant.) The implication is that race discrimination is more severe than indicated by the analysis that ignores the role of differences in the variances of the unobservables. Note that this evidence is consistent with a low level of standardization of X^{I*}, coupled with a higher estimated variance of the unobservable for blacks, as conjectured based on the earlier discussion of BM's results. In fact, as reported in the next row of the table, the estimated ratio of the standard deviation of the unobservable for blacks to the standard deviation for whites always exceeds one,

[34] And although not reported in the tables, this was true if the same methods used below to recover unbiased estimates of race discrimination were applied to the estimation of sex discrimination.

as reported in the next row of the table, although the difference is not statistically significant, paralleling the findings for the second part of the decomposition of the marginal effects.

The next two rows of the table report some diagnostic test statistics. First, the p-value from the test of the restrictions in equation (18) are shown, based on specifications interacting all of the controls with race. Recall that this restriction implies that the ratios of coefficients of blacks to whites are equal for all productivity-related characteristics, which is consistent with the differences between the coefficients for blacks and whites reflecting simply a difference in the variances of the unobservables. In all four cases this restriction is not rejected, with p-values ranging from 0.17 to 0.62. These test statistics imply that the restrictions necessary to recover an unbiased estimate of discrimination from the heteroskedastic probit model are not rejected. Nonetheless, the lower end of this range of p-values suggests that the restrictions sometimes might be fairly inconsistent with the data. As a consequence, below some alternative estimates are discussed that use only a subset of variables for which equation (18) is more consistent with the data.

Finally, the subset of control variables for which the absolute value of the estimated coefficient for whites exceeded that for blacks – consistent with the larger standard deviation of unobservables for blacks – was identified. Then, the heteroskedastic probit model was estimated leaving the race interactions of the *other* variables in the model, and the joint significance of these latter variables was tested. Despite this latter subset of variables having estimated coefficients less consistent with the restrictions in equation (18), the p-values indicate that these interactions can also be excluded from the model. This can be viewed as an overidentifying test of the restriction that there are no differences in the effects of any of the control variables by race, for the specifications for which the estimates are reported in the first row of Panel B. Although technically it is only necessary to assume that there is a single variable for which the coefficient is the same for blacks and whites, there is no obvious variable to choose for the purposes of identification; here, instead, I let the data select a set of variables more consistent with the identifying restriction.

Table 3 follows up on the last procedure, by instead simply dropping from the analysis the control variables for which the absolute value of the estimated coefficient for whites was less than for blacks. That is, only the control variables more consistent with the restrictions in equation (18) are retained. As we would expect, the p-values for the tests of this set of restrictions are now much closer to one, ranging from 0.68 to 0.92, compared with a range of 0.17 to 0.62 in Table 2. However, as the table shows, the estimated effects of race are similar to those in Table 2 and do not point to any different conclusions. That is not entirely surprising; given that the data come from a correspondence study in which control variables are randomly assigned, there is no reason that using a more parsimonious set of controls should have much impact on the estimated effects of race.

Monte Carlo Assessment of Estimation Procedure

This subsection provides evidence on how well the estimation procedure proposed in this paper work in terms of removing the bias in estimates of discrimination from correspondence study evidence. I first replicate results from Heckman (1998) illustrating the potential bias, and then show Monte Carlo results evaluating the performance of the heteroskedastic probit in eliminating this bias.

The ratio of the expressions in equations (14) and (14'), or

(21) $\Phi[(\beta_I X^{I*} + \gamma - c)/\sigma_B^{II}]/\Phi[(\beta_I X^{I*} - c)/\sigma_W^{II}]$,

is the relative hiring rate of blacks versus whites, for given values of the parameters and X^{I*}. The upper left-hand panel of Figure 1 replicates Figure 1 from Heckman (1998), in the case in which $c = 0$, $\beta_I = 1$, $Var(X_W^{II})/Var(X_B^{II}) = 2.25$,[35] and there is no discrimination ($\gamma = 0$). The dashed line, which matches Heckman's figure, is the analytical solution for the relative hiring rate at each value of X^{I*} (evaluated in steps of 0.1 over the range shown in the graph). The figure also has a solid line (barely distinguishable) that shows the mean estimate of the same ratio from Monte Carlo simulations. The assumed data generating process is $X^{I*} \sim N(0,1)$, $X_B^{II} \sim N(0,1)$, $X_W^{II} \sim N(0,2.25)$. The data are generated by sampling X^{I*} from a truncated normal distribution, in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$. The simulation is done 100 times at each value of X^{I*} shown in the graph, with samples of 2,000 blacks and 2,000 whites in each simulation (roughly BM's sample sizes). A probit model is estimated for each simulation, and the ratio in equation

[35] This is the ratio of the variances of the unobservables. Note that a larger value for whites is the opposite of the common assumption in models of statistical discrimination.

(21) is calculated; mean values of the estimated relative hiring rate at each value of X^{I*} are graphed.[36] The figure clearly illustrates the point that, despite the absence of discrimination in the data generating process, the evidence can either point to discrimination against blacks or discrimination in favor of blacks, depending on the level of standardization of X^{I*}; and the former evidence can, for the particular value of $Var(X_W^{II})/Var(X_B^{II})$ used in this example, appear very strong, with a relative hiring rate below 0.5 for low values of X^{I*}.[37]

The upper right-hand panel shows the mean estimates of γ – the coefficient on the black dummy variable – rather than the relative hiring rate, from these same simulations.[38] The same strong bias is obvious. Whereas the true value of γ equals zero, the mean estimate ranges from strongly negative to strongly positive. And, as we would expect, only at $X^{I*} = 0$ is the estimate of γ unbiased. The lower two panels of Figure 1 report the same kind of evidence, but in this case under the assumption of discrimination, with $\gamma = -0.5$. The same result is apparent, although of course the magnitudes differ.

To mimic the implementation of the estimation procedure described in this paper, which requires variation in the productivity-related characteristic X^1, I modify this procedure and generate data from two points of the distribution of X^{I*}. In particular, X^{I*} is now sampled from two truncated normal distributions, one using X^{I*}in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$, as before, and the second using instead $X^{I*} + 0.5$, again in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$. Figure 2 shows the same type of information as the right-hand panels in Figure 1, but now estimating the probit models for the simulated samples where – for each sample – data are drawn from two parts of the distribution of X^{I*}. The values shown on the horizontal axis are for the lower level of

---

36 Given that the probit estimates were very robust, only a small number of simulations was used.

37 There is perhaps a semantic issue about whether to interpret this behavior as discriminatory. It is true that in this case the productivity of blacks and whites are regarded equally by employers (or equivalently there is no taste discrimination). Moreover, employers are not making any assumption about mean differences in unobservables between blacks and whites. However, they are making assumptions about distributional differences with regard to the variance of unobservables, and it is these assumptions that lead them – given the level of standardization of the study applicants – to prefer one race over the other. To the best of my knowledge, such assumptions have never been viewed as discriminatory in either the economics or legal literature.

38 In this subsection the results are summarized in terms of estimates of γ rather than estimates of marginal effects. However, since the two marginal effects of interest, given in equations (19) and (20'), are multiplicative in the estimate of γ (in those equations the more generic β_k notation is used), results regarding biases in the estimates of γ carry over to the estimation of these marginal effects.

standardization; they range from −1.5 to 1, so the upper level of standardization ranges from −1 to 1.5.[39] The qualitative results are the same as in Figure 1. However, the biases in both the no discrimination and discrimination cases are a bit smaller than in Figure 1 because of the larger range covered by X^{I*}.[40]

Figure 3 reports results for the heteroskedastic probit estimation, using the simulations and data generating process used for Figure 2; in this case, though, we run 10,000 simulations for each pair of values of X^{I*}, as the heteroskedastic probit estimation was considerably less precise – as we would expect – than the simple probit estimation. The top panel covers the no discrimination case ($\gamma = 0$). The left-hand graph shows the true value of γ, as well as the mean and median of the simulations for each value of X^{I*}. These are largely indistinguishable in the figure, indicating no bias.[41] The right-hand panel provides evidence on the distribution of the estimates, showing the distance between the 25th and 75th percentiles of the estimates and between the 2.5th and 97.5th percentile at each value of X^{I*}. The distribution of estimates is quite tight at levels of standardization near the center of the distribution of X^{I*}, but becomes wider at more extreme values, when hiring rates in the generated data move towards zero or one.

The discrimination case ($\gamma = -0.5$) is a bit different, with some evidence of bias towards zero – or in other words bias *against* finding evidence of discrimination. Thus, the heteroskedastic probit estimation, in samples of the size available in the BM study (and likely in the ballpark of feasible sample sizes for correspondence studies) will, if anything, be conservative in finding evidence of discrimination.

The last analysis, reported in Figure 4, considers the implications of the data generating process violating the identifying assumption that the coefficient(s) on the productivity-related characteristics are

[39] Given that $X^{I*} \sim N(0,1)$, the lower and upper ends of this range imply choosing very low-quality and very high-quality applicants. Presumably in a correspondence study one would not want to, for example, construct applicants at something like two standard deviations below the mean (i.e., 2.5th percentile of the distribution).

[40] Note also that in this case the unbiased estimate occurs at the value of −0.25 (for X^{I*}) on the horizontal axis, where the average of the upper and lower standardization levels equals zero. The reduction in bias is little less clear in the discrimination case. To clarify, the bias in Figures 1 and 2 should be contrasted at comparable value of X^{I*}, given that Figure 2 shows the mean estimates at the lower level of standardization of X^{I*}. For example, for the discrimination case, the mean estimate of γ at $X^{I*} = 1$ in Figure 1 should be compared to the mean estimate at $X^{I*} = 0.75$ in Figure 2 (in which case this is the lower standardization level and the average is 1); the latter estimate is in fact closer to zero.

[41] A more detailed examination of the estimates shows no bias in the median of the estimates over the range shown in the graph, whereas the mean estimates are responsible for the slight deviations at the upper and lower ranges of X^{I*}, with perhaps a slight tendency to underestimate γ for low values of X^{I*}, and vice versa.

equal for blacks and whites.[42] Results are presented for two cases: mild violation in which the coefficient on X^{I*} (β_I) is slightly larger for whites than for blacks (1.1 versus 1); and strong violation in which it is much larger (2 versus 1). As Figure 4 shows, in the case of no discrimination the results are indistinguishable from when the identifying assumption is not violated. In contrast, in the discrimination case, comparing the right-hand panels in Figure 4 to the bottom left-hand panel of Figure 3, we see that the estimate of γ becomes more negative, changing only slightly and remaining (in absolute value) below the true value of −0.5 in the first case, but rising above it in the second case.[43]

Thinking about how γ is identified, these findings make sense. Using estimates of the separate probits in equations (14) and (14'), the ratio of the standardized white probit coefficient (setting $\sigma_W^{II} = 1$) to the black probit coefficient (β_I/σ_B^{II}) identifies σ_{BR}^{II} (which, recall, equals $\sigma_B^{II}/\sigma_W^{II}$). When the true value of β_I is larger for whites than for blacks, but it is assumed that they are equal, σ_{BR}^{II} is overestimated.[44] Recall from the earlier discussion that the probit for blacks identifies $(-c + \gamma)/\exp(\omega) = (-c + \gamma)/\sigma_{BR}^{II}$. Because c = 0 in the simulations, we identify γ by multiplying the estimate of this expression by the estimate of σ_{BR}^{II}; the upward bias in the estimate of σ_{BR}^{II} therefore implies that the estimate of γ is biased away from zero. In the no discrimination case, when $\gamma = 0$, this is irrelevant; multiplying an estimate that averages zero by the upward-biased estimate of σ_{BR}^{II} has no effect. But when the true γ is non-zero (and negative), this bias leads to an estimate of γ that is more negative. Of course this latter result depends on the direction in which the identifying assumption is violated. Clearly if the analysis had been done with a violation of the assumption in the opposite direction (β_I larger for blacks), then the estimate of γ would be biased towards zero. Nonetheless, it follows from this reasoning that the bias is multiplicative, and hence does not generate the wrong *sign* for the estimate of γ.

This Monte Carlo evaluation leads to a few conclusions. First, when there is in fact no discrimination, the heteroskedastic probit estimation leads to an unbiased estimate of discrimination,

[42] Recall that this assumption must hold true for at least some selected set of productivity-related characteristics to identify the ratio of the variance of the unobservables for the two groups.

[43] Although not shown in the graph, for the analyses in Figure 4 there was substantial deterioration in the estimation of γ for X^{I*} somewhat further to the left of the range shown in the graph.

[44] For example, in the case in the top panel of Figure 4, the ratio of coefficients is $(\beta_I \cdot 1.1)/(\beta_I/\sigma_{BR}^{II}) = 1.1 \cdot \sigma_{BR}^{II}$.

correcting the problem of the wildly varying (and generally biased) estimates that could otherwise result from the analysis of data from a correspondence study. Second, when there is discrimination, there is a bias against finding evidence of discrimination (when the identifying assumption holds). Nonetheless, the bias is generally constant for a reasonable range of the productivity-related characteristics of applicants, in which case the conclusions will not vary with the level of standardization and the relative variances of the unobservables. Again, then, the heteroskedastic probit procedure eliminates the problem that correspondence study evidence can lead to different answers depending on the level of standardization and the relative variances of the unobservables, with no change in underlying discriminatory behavior. Moreover, the estimator does not generate evidence of discrimination against one group when there is in fact discrimination against the other group, which can happen with standard probit analysis of data from a correspondence study. However, it is possible that we may still fail to detect statistically significant evidence of discrimination when it exists.

Finally, the procedure requires an identifying assumption that the true probit coefficients for at least one productivity-related characteristic are equal for the two groups studied. When this is violated, in the case of no discrimination the procedure still appears to provide unbiased evidence. But in the case where there is discrimination, additional bias in the estimate of γ is introduced, which could either exacerbate or offset the bias noted above. It is reassuring, though, that apparently even in this case the estimation procedure proposed in this paper will still not be biased in the direction of finding evidence of discrimination (in either direction) when there is in fact no discrimination, nor will it give evidence of discrimination against one group when there is discrimination against the other group.

VI. Conclusions and Discussion

Many researchers view audit and correspondence studies as the most compelling way to test for labor market discrimination. And evidence from applying these methods to many different types of groups tends to find evidence of discrimination. The use of audit studies to test for labor market discrimination has been criticized on numerous grounds having to do with whether applicants from different groups appear identical to employers. Many of these criticisms can be countered by using correspondence studies in which fictitious applicants on paper are substituted for fictitious applicants in person.

However, Heckman and Siegelman (1993) show that even in correspondence studies in which group averages are identical conditional on the controls, group differences in the variances of unobservable dimensions of productivity can invalidate the empirical tests, leading to spurious evidence of discrimination, or spurious evidence of an absence of discrimination. This is an important criticism of correspondence studies, as it implies that evidence regarding discrimination from even the best-designed correspondence study can give misleading evidence of discrimination – or of its absence. Nonetheless, the criticism has been ignored in the literature.

This paper shows that if the correspondence study includes observable measures of variation in applicants' quality that affect hiring outcomes, an unbiased estimate of discrimination can be recovered even when there are group differences in the variances of the unobservables. The method is applied to Bertrand and Mullainathan's (2004) correspondence study, and leads to stronger evidence of race discrimination that adversely affects blacks than the evidence that results when differences in the variances of the unobservables are ignored. Moreover, this conclusion is bolstered by Monte Carlo simulations suggesting that the estimation procedure performs well, essentially eliminating the problems highlighted by Heckman and Siegelman that could otherwise lead to badly misleading conclusions from the analysis of data from correspondence (or audit) studies.

Finally, it should be recognized that the method proposed here can be easily implemented in any future correspondence study. All that is needed is for the resumes or applications to include a set of characteristics that are linked to variation in the probability of being hired. This is different from what is sometimes done in designing correspondence studies, where researchers try to create a bank of resumes of nearly equally-qualified candidates. Clearly researchers sometimes inadvertently create resumes of different quality. All that needs to be done, however, is to intentionally do this. Once a researcher confirms that a set of resume characteristics affected hiring outcomes, it is then possible to test for bias from different variances of the unobservables for the two groups, and to obtain estimates free of this bias.

References

Aigner, Dennis J., and Glen Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review*, Vol. 30, No. 2, January, pp. 175-87.

Banerjee, Abhijit, Marianne Bertrand, Saugato Datta, and Sendhil Mullainathan. 2008. "Labor Market Discrimination in Delhi: Evidence from a Field Experiment." *Journal of Comparative Economics*, Vol. 38, No. 1, March, pp. 14-27.

Becker, Gary S. 1971. The Economics of Discrimination, Second Edition. Chicago: University of Chicago Press.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, Vol. 94, No. 4, September, pp. 991-1013.

Bovenkerk, F., M. Gras, and D. Ramsoedh. 1995. "Discrimination Against Migrant Workers and Ethnic Minorities in Access to Employment in the Netherlands." *International Migration Papers*, No. 4, International Labour Office, Geneva, Switzerland.

Cornelißen, Thomas. 2005. "Standard Errors of Marginal Effects in the Heteroskedastic Probit Model." Institute of Quantitative Economic Research, Discussion Paper No. 230.

Darity, William A., Jr., and Patrick L. Mason. 1998. "Evidence on Discrimination in Employment: Codes of Color, Codes of Gender." *Journal of Economic Perspectives*, Vol. 12, No. 2, Spring, pp. 63-90.

Fix, Michael, and Raymond Struyk. 1993. Clear and Convincing Evidence: Measurement of Discrimination in America. Washington, DC: The Urban Institute Press.

Foster, Andrew, and Mark Rosenzweig. 1993. "Information, Learning, and Wage Rates in Low-Income Rural Areas." *Journal of Human Resources*, Vol. 28, No. 4, Fall, pp. 759-790.

Heckman, James J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives*, Vol. 12, No. 2, Spring, pp. 101-16.

Heckman, James, and Peter Siegelman. 1993. "The Urban Institute Audit Studies: Their Methods and Findings." In Fix and Struyk, eds., Clear and Convincing Evidence: Measurement of Discrimination in America. Washington, DC: The Urban Institute Press, pp. 187-258.

Hellerstein, Judith K., and David Neumark. 1999. "Sex, Wages, and Productivity: An Empirical Analysis of Israeli Firm-Level Data." *International Economic Review*, Vol. 40, No. 1, February, pp. 95-123.

Hellerstein, Judith, and David Neumark. 2006. "Using Matched Employer-Employee Data to Study Labor Market Discrimination." In William Rodgers, ed., Handbook on the Economics of Discrimination. Great Britain: Edgar Elgar Publishing, pp. 29-60.

Johnson, Norman L., and Samuel Kotz. 1970. Continuous Univariate Distributions – 2. New York: John Wiley and Sons.

Lundberg, Shelly J., and Richard Startz. 1983. "Private Discrimination and Social Intervention in Competitive Labor Markets." *American Economic Review*, Vol. 73, No. 3, June, pp. 340-7.

Maddala, G.S. 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.

Mincy, Ronald. 1993. “The Urban Institute Audit Studies: Their Research and Policy Context.” In Fix and Struyk, eds., Clear and Convincing Evidence: Measurement of Discrimination in America. Washington, DC: The Urban Institute Press, pp. 165-86.

Neumark, David. 1996. “Sex Discrimination in Restaurant Hiring: An Audit Study.” *Quarterly Journal of Economics*, Vol. 111, No. 3, August, pp. 915-41.

Neumark, David. 1999. “Wage Differentials by Race and Sex: The Roles of Taste Discrimination and Labor Market Information.” *Industrial Relations*, Vol. 38, No. 3, July, pp. 414-45

Oaxaca, Ronald. 1973. “Male-Female Wage Differentials in Urban Labor Markets.” *International Economic Review*, Vol. 14, No. 3, October, pp. 693-709.

Pager, Devah. 2007. “The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future.” *The Annals of the American Academy of Political and Social Science*, Vol. 609, No. 1, pp. 104-33.

Riach, Peter A., and Judith Rich, 2002. “Field Experiments of Discrimination in the Market Place.” *The Economic Journal*, Vol. 112, No. 483, November, pp. F480-518.

Smeeters, B., and A. Nayer. 1998. “La Discrimination a l’Acces a l’Emploi en Raison de l’Origine Etrangere: le Cas de le Belgique.” *International Migration Papers*, No. 23, International Labour Office, Geneva, Switzerland.

Turner, Margery, Michael Fix, and Raymond Struyk. 1991. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring*. UI Report 91-9. The Urban Institute, Washington, DC.

U.S. Equal Employment Opportunity Commission. 1996. Notice 915.002, May 22, http://www.eeoc.gov/policy/docs/testers.html (viewed February 28. 2010).

Williams, Richard. 2009. “Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups.” Unpublished manuscript, Department of Sociology, Notre Dame University.

Table 1
Employment Probits for Callbacks: Basic Results

	Males and females			Females		
	(1)	(2)	(3)	(4)	(5)	(6)
Black	-.033	-.030	-.030	-.033	-.030	-.030
	(.006)	(.006)	(.006)	(.008)	(.007)	(.007)
Female	.009	-.001	.001	…	…	…
	(.012)	(.011)	(.011)			
Selected individual resume controls						
Bachelor's degree		.009	.009		.019	.019
		(.009)	(.009)		(.010)	(.010)
Experience $\cdot 10^{-1}$		.080	.076		.080	.076
		(.029)	(.028)		(.034)	(.033)
Experience^2 $\cdot 10^{-2}$		-.022	-.021		-.019	-.018
		(.011)	(.010)		(.013)	(.012)
Academic honors		.039	.040		.026	028
		(.015)	(.015)		(.017)	(.017)
Special skills		.056	.055		.060	.059
		(.009)	(.009)		(.010)	(.010)
Other controls:						
Individual resume characteristics		X	X		X	X
Neighborhood characteristics			X			X
Mean callback rate	.080	.080	.080	.082	.082	.082
N	4,784	4,784	4,784	3,670	3,670	3,670

Notes: Marginal effects using equation (19) are reported. Standard errors are computed clustering on the ad to which the applicants responded, and are reported in parentheses; the delta method is used to compute standard errors for the marginal effects. Individual resume characteristics include bachelor's degree, experience and its square, volunteer activities, military service, having an email address, gaps in employment history, work during school, academic honors, computer skills, and other special skills. Neighborhood characteristics include the fraction high school dropout, college graduate, black, and white, as well as log median household income, in the applicant's zip code.

Table 2
Employment Probits for Callbacks: Fully Interactive Specifications, Test Statistics, Heteroskedastic Models, and Comparisons with Standard Estimates

	Males and females		Females	
	(1)	(2)	(3)	(4)
A. Estimates from basic probit (Table 1)				
Black	-.030	-.030	-.030	-.030
	(.006)	(.006)	(.007)	(.007)
B. Heteroskedastic probit model				
Black (unbiased estimates)	-.024	-.026	-.026	-.027
	(.007)	(.007)	(.008)	(.008)
Effect of race through level	-.086	-.070	-.072	-.054
	(.038)	(.040)	(.040)	(.040)
Effect of race through variance	.062	.045	.046	.028
	(.042)	(.043)	(.045)	(.044)
Standard deviation of unobservables, black/white	1.37	1.26	1.26	1.15
Wald test statistic, null hypothesis that ratio of standard deviations = 1 (p-value)	.22	.37	.37	.56
Wald test statistic, null hypothesis that ratios of coefficients for whites relative to blacks are constant, fully interactive probit model (p-value)	.62	.42	.17	.35
Test overidentifying restrictions: include in heteroskedastic probit model interactions for variables with |white coefficient| < |black coefficient|, Wald test for joint significance of interactions (p-value)	.83	.33	.34	.56
Number of overidentifying restrictions	3	6	2	6
Other controls:				
Individual resume characteristics	X	X	X	X
Neighborhood characteristics		X		X
N	4,784	4,784	3,670	3,670

Notes: See notes to Table 1. In the first row of Panel B the marginal effects in equation (20) are reported, with the decomposition in equations (20') and (20'') immediately below. The standard errors for the two components of the marginal effects are computed using the delta method. Test statistics are based on the variance-covariance matrix clustering on the ad to which the applicants responded. Individual resume characteristics also include the variables listed separately in Table 1.

Table 3
Employment Probits for Callbacks: Restricted Specifications Using only Controls with Absolute Value of Estimated Effect Larger for Whites than Blacks in Fully Interactive Probit Model

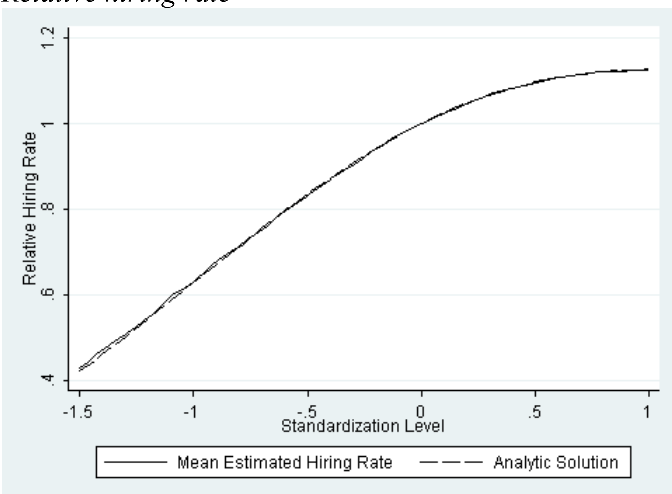
	Males and females		Females	
	(1)	(2)	(3)	(4)
A. Estimates from basic probit				
Black	-.030	-.030	-.030	-.030
	(.006)	(.006)	(.007)	(.006)
B. Heteroskedastic probit model				
Black (unbiased estimates)	-.024	-.025	-.024	-.025
	(.007)	(.007)	(.009)	(.008)
Effect of race through level	-.090	-.080	-.086	-.077
	(.037)	(.036)	(.040)	(.038)
Effect of race through variance	.066	.056	.062	.052
	(.041)	(.039)	(.044)	(.042)
Standard deviation of unobservables, black/white	1.41	1.33	1.37	1.30
Wald test statistic, null hypothesis that ratio of standard deviations = 1 (p-value)	.19	.23	.25	.29
Wald test statistic, null hypothesis that ratios of coefficients for whites relative to blacks are constant, fully interactive probit model (p-value)	.84	.92	.68	.74
Other controls:				
Individual resume characteristics	X	X	X	X
Neighborhood characteristics		X		X
N	4,784	4,784	3,670	3,670

Notes: See notes to Tables 1 and 2.

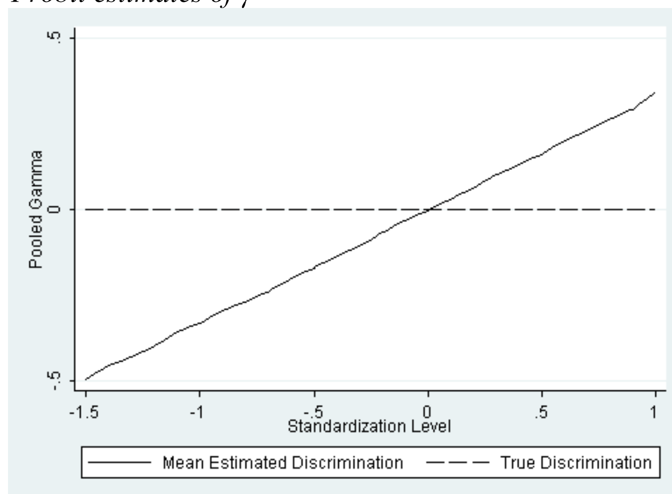
Figure 1
Replication of Heckman (Figure 1, 1998), and Monte Carlo Simulations of Simple Probit Estimation

No discrimination ($\gamma = 0$)

Relative hiring rate



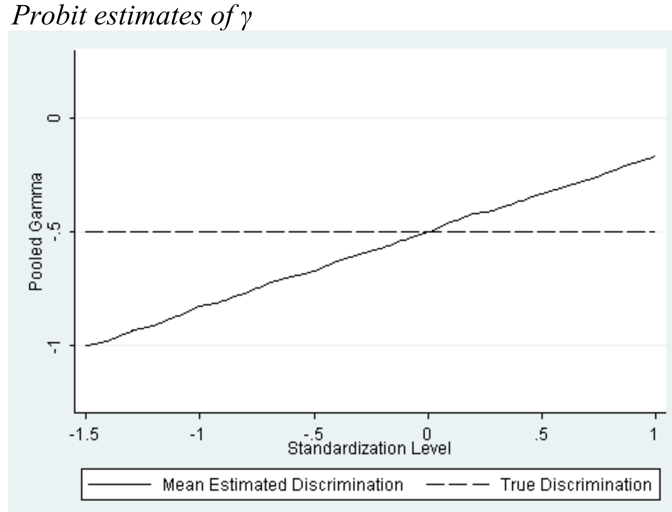
Probit estimates of γ



Discrimination ($\gamma = -.5$)

Relative hiring rate

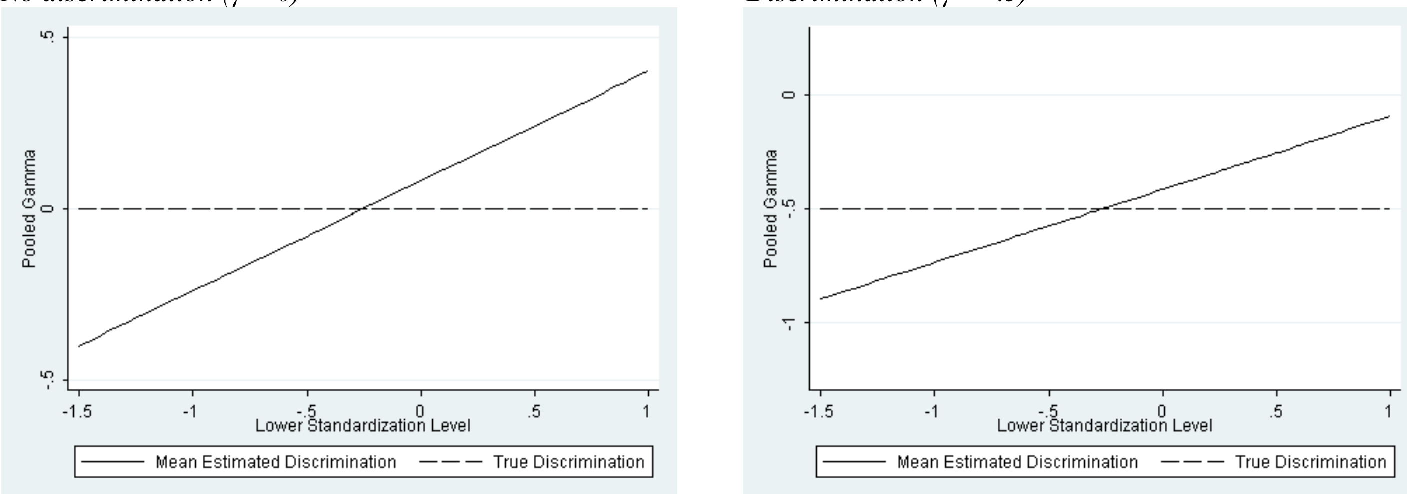
Probit estimates of γ



Notes: Left-hand graph shows relative hiring rate equal to ratio of equation (14) to equation (14') = $\Phi[(\beta_I X^{I*} + \gamma - c)/\sigma_B^{II}]/\Phi[(\beta_I X^{I*} - c)/\sigma_W^{II}]$. Right-hand graph shows mean estimates of γ. In the data generating process, $X^{I*} \sim N(0,1)$, $X_B^{II} \sim N(0,1)$, $X_W^{II} \sim N(0,2.25)$, so $Var(X_W^{II})/Var(X_B^{II}) = 2.25$ (X_W^{II} and X_B^{II} are unobservable); $\beta_I = 1$ and $c = 0$ for both blacks and whites. In left-hand graph, dashed line is analytic solution. Solid line is generated by Monte Carlo simulation, drawing 4,000 observations (2,000 white and 2,000 black) from truncated normal distribution at each value of X^{I*} (in steps of $0.1 \pm 0.1 \cdot SD(X^{I*})$) and estimating probit model. Simulation is done 100 times at each value of X^{I*} shown in graph.

Figure 2
Monte Carlo Simulations of Simple Probit Estimation, Bias in Estimate of Discrimination with Two Types of Applicants

No discrimination (γ = 0) *Discrimination (γ = -.5)*

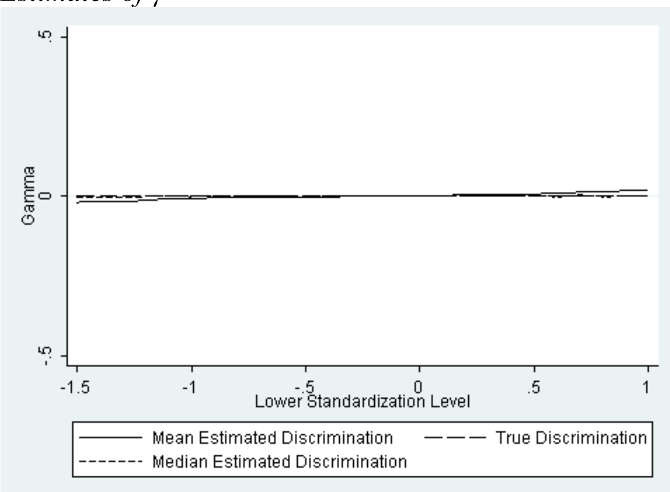


Notes: Estimates are generated by Monte Carlo simulation, drawing 4,000 observations (2,000 white and 2,000 black) observations from two truncated normal distributions (one at each value of X^{I*} (in steps of 0.1 $\pm$ 0.1·SD(X^{I*})), and one at each value of X^{I*} + .5 (again in steps of 0.1) $\pm$ 0.1·SD(X^{I*})), and estimating probit model. Simulation is done 100 times at each value of X^{I*} shown in graph. As in Figure 1, the data generating process has $X^{I*} \sim N(0,1)$, $X_B^{II} \sim N(0,1)$, $X_W^{II} \sim N(0,2.25)$, so $Var(X_W^{II})/Var(X_B^{II}) = 2.25$ (X_W^{II} and X_B^{II} are unobservable); and $\beta_I = 1$ and c = 0 for both blacks and whites. Graphs show mean estimates of γ from probit estimation.

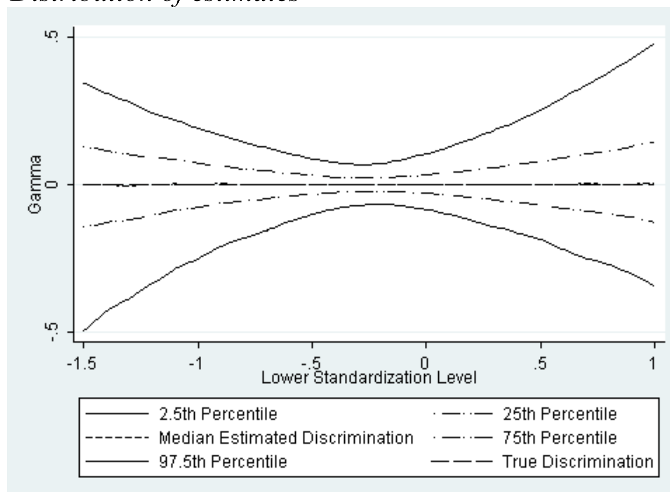
Figure 3
Monte Carlo Simulations of Heteroskedastic Probit Estimation, Estimates of γ and Distributions

No discrimination (γ = 0)

Estimates of γ

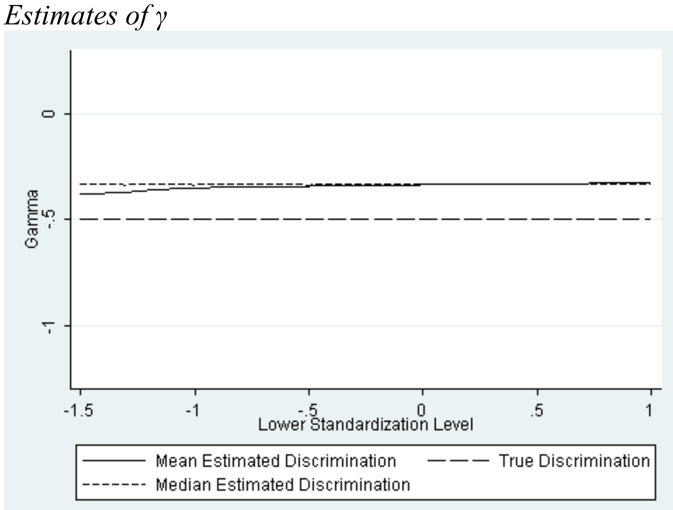


Distribution of estimates

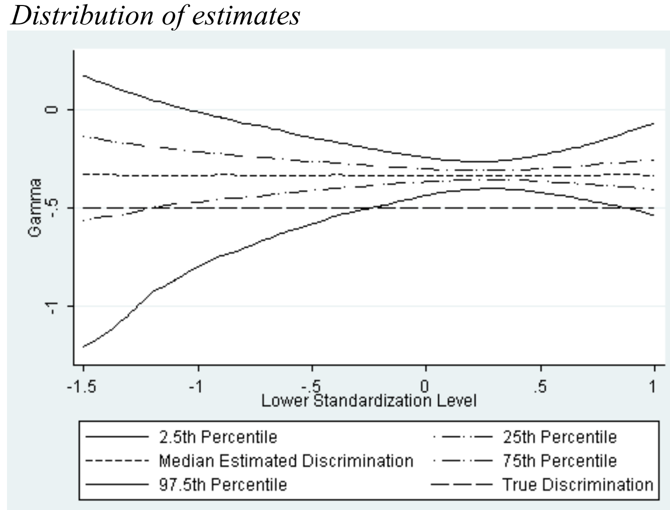


Discrimination (γ = -.5)

Estimates of γ



Distribution of estimates

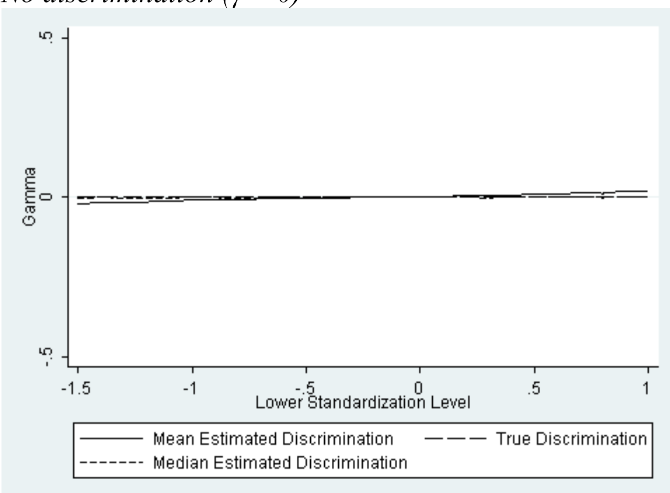


Notes: Estimates are generated by Monte Carlo simulation, drawing 4,000 observations (2,000 white and 2,000 black) observations from two truncated normal distributions (one at each value of X^{I*} (in steps of 0.1 $\pm$ 0.1·SD(X^{I*})), and one at each value of X^{I*} + .5 (again in steps of 0.1 $\pm$ 0.1·SD(X^{I*})), and estimating heteroskedastic probit model. Simulation is done 10,000 times at each value of X^{I*} shown in graph. As in Figure 1, the data generating process has $X^{I*} \sim N(0,1)$, $X_B^{II} \sim N(0,1)$, $X_W^{II} \sim N(0,2.25)$, so $Var(X_W^{II})/Var(X_B^{II}) = 2.25$ (X_W^{II} and X_B^{II} are unobservable); and $\beta_I = 1$ and $c = 0$ for both blacks and whites. Left-hand graphs show mean and median estimates of γ from heteroskedastic probit estimation, and right-hand graphs show distributions of estimates.

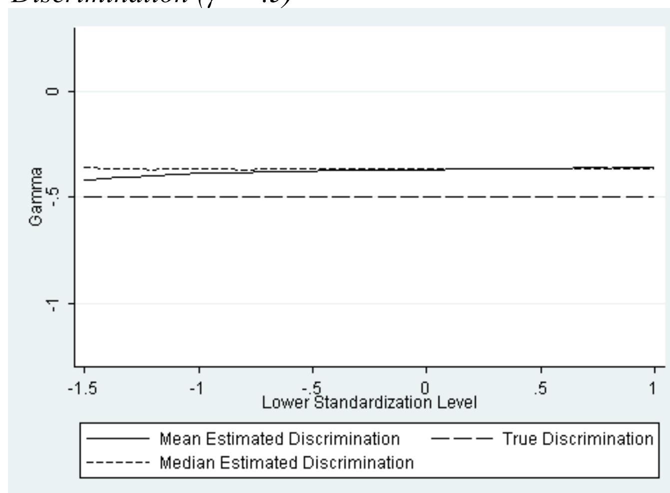
Figure 4
Monte Carlo Simulations of Heteroskedastic Probit Estimation, with Model Misspecification Masking Higher Unobserved Variance for Whites, Estimates of γ

Mild violation of identifying assumption in data generating process (β_1 for whites = 1.1)

No discrimination (γ = 0)

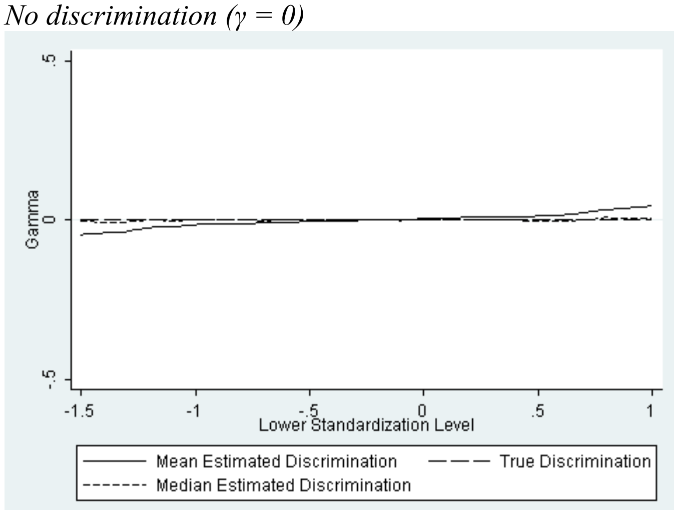


Discrimination (γ = -.5)

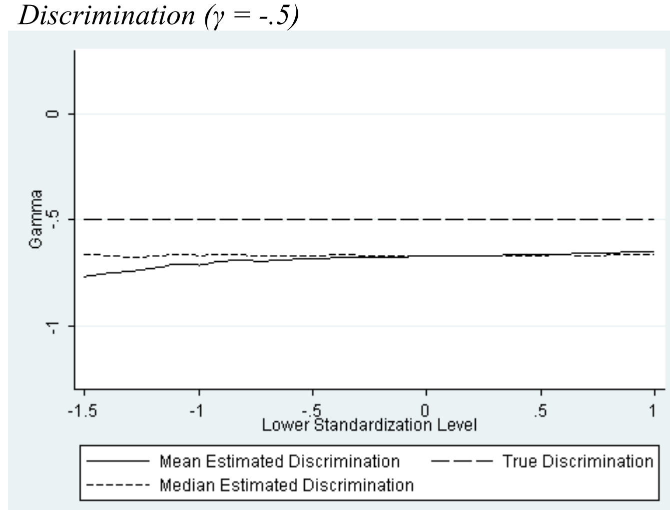


Strong violation of identifying assumption in data generating process (β_1 for whites = 2)

No discrimination (γ = 0)



Discrimination (γ = -.5)



Notes: See notes to Figure 3. The only differences are that β_1 is unequal for blacks and whites; it is always equal to 1 for blacks, and as indicated in the graph headings for whites.