NBER WORKING PAPER SERIES

DO FIELD EXPERIMENTS ON LABOR AND HOUSING MARKETS OVERSTATE DISCRIMINATION? RE-EXAMINATION OF THE EVIDENCE

David Neumark
Judith Rich

Working Paper 22278
http://www.nber.org/papers/w22278

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2016

We wish to thank the following authors of studies who generously provided their raw data: Ali Ahmed, Lina Andersson, and Mats Hammarstedt; Stijn Baert, Bart Cockx, Niels Gheyle, and Cora Vandamme; Marianne Bertrand and Sendhil Mullainathan; Mariano Bosch, M. Angeles Carnero, and Lídia Farré; Magnus Carlsson and Stefan Eriksson; Dan-Olof Rooth (also with Magnus Carlsson); Nick Drydakis; Michael Ewens, Bryan Tomlin, and Liang Choon Wang; and Phil Oreopoulos. We thank Nick Drydakis and Philip Oreopoulos for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Do Field Experiments on Labor and Housing Markets Overstate Discrimination? Re-examination of the Evidence
David Neumark and Judith Rich
NBER Working Paper No. 22278
May 2016
JEL No. J71

ABSTRACT

There have been over 60 field experiments on discrimination in labor and housing markets conducted since 2000, in 16 countries. These studies nearly always find significant levels of discrimination against minority transactors in these markets. A key challenge to these findings, though, is that even in rather ideal conditions, the estimates of discrimination can be biased if there is differential variation in the unobservable determinants of productivity of majority and minority groups, conditional on the characteristics of market participants these experiments reveal to employers or landlords (Heckman, 1998). The potential bias could go in either direction, but naturally raises the question of whether this experimental literature as a whole overstates the evidence of discrimination. To assess this question, we re-assess the evidence from the nine existing studies that have sufficient information to implement a correction for this bias (Neumark, 2012). For the housing market studies, the estimated effect of discrimination is robust to this correction. For the labor market studies, in contrast, the evidence is less robust; in about half of cases covered in these studies, the estimated effect of discrimination either falls to near zero or becomes statistically insignificant.

David Neumark
Department of Economics
University of California at Irvine
3151 Social Science Plaza
Irvine, CA 92697
and NBER
dneumark@uci.edu

Judith Rich
Economics and Finance Department
University of Portsmouth
Richmond Building Portland Street
Portsmouth PO1 3DE
United Kingdom
judy.rich@port.ac.uk

1. Introduction

Field experiments – specifically, audit or correspondence studies – have been used extensively to test for discrimination in markets. In audit studies of labor market discrimination, fake job candidates ("testers") of different races, ethnicities, etc., who are sometimes actors, are sent to interview for jobs (or in some early studies, apply by telephone). The candidates have similar resumes and are often trained to act, speak, and dress similarly. Correspondence studies, in contrast, use fictitious job applicants who exist on paper only (or now, electronically), and differ systematically only on group membership. The response captured in correspondence studies is a "call-back" for an interview or a closely related positive response. In contrast, the final outcome in audit studies is actual job offers. Differences in outcomes between groups are likely attributable to discrimination, although there are, naturally, some subtle issues of interpretation. Audit and correspondence (AC) studies have also been used to study discrimination in housing markets and consumer markets. AC studies are widely regarded as providing more rigorous evidence on discrimination than can be obtained from non-experimental evidence in which group membership may be correlated with unobservables.[1]

Nonetheless, AC studies have come in for criticism (Heckman and Siegelman, 1993; Heckman, 1998). The most challenging criticism of these studies is that, in the standard implementation, the resulting estimate of discrimination can be biased in either direction – or equivalently, discrimination can be unidentified. This problem arises when the variances of the unobservables differ across the groups studied, something that cannot be ruled out or easily controlled in AC studies, and indeed is at the core of early models of statistical discrimination (Aigner and Cain, 1977). This criticism – which we refer to (perhaps unfairly to Siegelman) as the

[1] The methods and empirical findings from these studies have been reviewed by Pager (2007), Riach and Rich (2002), Rich (2014), and Neumark (2016).

"Heckman critique" – holds even under quite ideal conditions (detailed later) in which other potential research design flaws that Heckman and Siegelman discuss are absent.

A statistical method that can lead to unbiased estimates of discrimination using data from AC studies was proposed in Neumark (2012). As explained below, most past AC studies do not have the requisite data – applicant or other characteristics aside from the identifier for the group in question that shift the probability of call-backs or hires. However, we have identified nine studies of discrimination against minorities (based on race, ethnicity, or sexual orientation) in labor and housing markets conducted over the last couple of decades that do include the requisite data.[2]

These nine studies – just like nearly all of the far greater number of AC studies that do *not* have the requisite data – find evidence of discrimination against ethnic or racial minorities, immigrants, or gays and lesbians.[3] However, we have obtained the original data from the authors. Our goal in this paper is to test whether this evidence is robust to confronting the data with the Heckman critique. Specifically, implementing the correction for bias from differences in the variances of unobservables across groups, do these studies still uniformly point to discrimination?

After providing some background details on these studies, we explain the approach and report results. To summarize the results briefly, for the housing market studies the estimated effects of discrimination are robust to this correction. For the labor market studies, in contrast, the evidence is less robust; in about half of the cases covered in these studies, the estimated effect of discrimination either falls to near zero or becomes statistically insignificant. The results for the labor market, in particular, suggest that researchers need to build into future AC studies the data and experimental design needed to address the Heckman critique, and that further work on

---

[2] The studies are: Ahmed et al. (2010); Baert et al. (2013); Bertrand and Mullainathan (2004) – the data used in Neumark (2012); Bosch et al. (2010); Carlsson and Eriksson (2014); Carlsson and Rooth (2007); Drydakis (2014); Ewens et al. (2014); and Oreopoulos (2011).

[3] For the most recent review of a large number of AC studies, see Neumark (2016).

different ways to eliminate bias from AC studies estimates of discrimination is warranted. More substantively, our re-examination of the evidence suggests that the overall body of experimental evidence on labor market discrimination provides a less clear signal of discrimination than one would draw from the results reported in the existing studies.

2. The field experiments covered in this paper

The field experiments re-analyzed in this paper are one of three broad types: studies of ethnic/immigrant or race discrimination in labor markets; studies of sexual orientation discrimination in labor markets; and studies of ethnic/immigrant or race discrimination in rental housing markets. Many of the details and results of these studies are discussed in Rich (2014) and Neumark (2016). Here we focus only on what is essential to understand the analysis of bias from differences in unobservables that we implement in this paper. Readers interested in more details on these specific studies, and the techniques used more generally, should see these surveys (or of course the original papers). We do not go into more detail because our goal in this paper is not to compare or critique other dimensions of these studies, but rather just to consider the robustness of the conclusions to addressing the Heckman critique.

What distinguishes these nine studies from the others in the literature is that they use applicants distinguished not only by race, ethnicity (including immigrant origin), or sexual orientation, but also by different levels of qualifications. In these studies, this was done to ask, in a general way, whether the evidence of discrimination by ethnicity, race, or sexual orientation differed for applicants with different levels of qualifications.[4] As discussed in the next section,

[4] The first study of this type (Jowell and Prescott-Clarke, 1970) considered this issue. The study compared job offer outcomes for immigrant versus white British applicants, and gave half the applications in each group higher qualifications with regard to education. (There was also variation among the immigrants only in whether they were English-speaking and whether secondary education was in Britain, although this kind of variation that does not apply equally to majority and minority groups is not as useful.) The more recent

however, the availability of data with variation in applicant qualifications is exactly what is needed to implement the empirical method that addresses the Heckman critique.

Baert et al. (2015), Bertrand and Mullainathan (2004), Carlsson and Rooth (2007), and Drydakis (2014) all used matched pairs of applicants, with two applications sent to each job vacancy. Oreopoulos (2011) considered differences for many different ethnic groups (relative to native Canadians), in some cases also signaling immigrant status, and sent multiple resumes for each job vacancy. Across these studies, on the resumes used, which were either real resumes the authors found or resumes generated randomly from elements of other resumes, race or ethnicity was signalled by name, and immigrant status in addition to ethnicity was sometimes further signalled by education or work experience in a foreign country (Oreopoulos, 2011). Sexual orientation was signalled by participation in an organization active on behalf of the gay community or a gay organization.

There have been fewer studies of discrimination in housing markets in the broader literature. In the housing market experiments we re-examine, only Bosch et al. (2010) used matched pairs, while the other three (Ahmed et al., 2010; Ewens et al., 2014; Carlsson and Eriksson, 2014) sent a single rental enquiry. An accompanying message providing details on the applicant was attached, in which the researchers manipulated the information provided – ethnicity and race, as well as other qualifications. In these studies, signaling is done by name, although Bosch et al. (2010) interpret their results for Moroccan versus Spanish names as measuring discrimination against immigrants.

Other qualifications also varied across the resumes. For example, Bertrand and Mullainathan (2004) generally sent four applications to each job. They created two matched pairs of applicants, one with low-quality background and another pair with high-quality background.

---

studies with such data that we re-examine in the present paper are those for which we could recover the data from authors.

The quality of the applicant was varied using labor market experience, career profiles, employment history, and skills such as employment experience gained either over summer or while at school, volunteering, extra computer skills, certification degrees, foreign language skills, honors, or some military experience (Bertrand and Mullainathan, 2004, pp. 994-5). Carlsson and Rooth (2007) signalled similar additional information as Bertrand and Mullainathan on applicants as well as different spells of unemployment, work experience over summer, overqualified or not, personality traits, and cultural and sporting activities listed as hobbies and interests. Oreopoulos (2011) varied the information provided on the extent of foreign education and foreign experience as well as language skills and certification and masters degrees. Drydakis (2014) used an accompanying cover letter to provide different types of information. In the housing market tests, researchers manipulated the information on the applicant using an accompanying message to explore the impact of basic, negative, or positive information.

The richness and number of qualifications that researchers chose to vary across the applicants differ quite a bit across these studies. For the labor market studies, these qualifications generally pertain to education, experience, and skills, but sometimes extend to attempts to convey something about the applicant's personality or hobbies, the order of the application, and other things. One of the housing studies (Carlsson and Eriksson, 2014) tries to provide information on the applicant's lifestyle, which could be relevant to a potential landlord. We do not discuss the different qualifications used in each study in detail, but list them for each study in the tables reporting the statistical analysis (Tables 2A and 2B for the labor market studies, and Table 3 for the housing market studies). The reader will note that we also list other features of the ads that could affect the probability of a call-back – such as characteristics of the job or the apartment. We include these because – as explained in the next section – the statistical method is informed by differences in the coefficients between the two groups studied in *any* of the factors that can affect call-backs.

3. Findings from the field experiments covered in this paper

Table 1 summarizes the results from the nine studies we re-examine, as well as giving basic information about them, including the years covered, the groups covered, and the outcomes. The original studies report results in different ways, varying between chi-square/Fisher exact tests, binomial tests, or tests of the null hypothesis that there is no difference in the call-back rate between the groups, typically controlling for other aspects of the resumes. However, here we report results on a consistent basis for all studies – marginal effects from probit models using the full set of resume characteristics included in the data – which we have estimated from data provided by the authors of these studies.[5]

As reported in Table 1, the five labor market experiments covered in Panel A all find statistically significant evidence of discrimination against either ethnic minorities, blacks, or gays and lesbians. The estimated differentials by racial and ethnic groups are in the same range – an approximately 0.03 to 0.10 lower probability of a call-back. These are on somewhat different baseline rates of call-backs, but the call-back rates also do not vary that much across these studies.[6] The two estimates from Drydakis (2014), for discrimination against gays and lesbians in Cyprus, are much larger (although the baseline call-back rates are much higher too).

The four housing market studies similarly find consistent evidence of discrimination against minorities. The range of estimates is fairly tight (a 0.09 to 0.17 lower call-back rate). Thus, every one of these studies points to evidence of discrimination against the minority group.

The conclusions from these studies strongly echo the broader literature, in which nearly

---

[5] Details on the control variables, the standard errors, etc., are provided in tables discussed below. Not surprisingly, the results in Table 1 closely parallel the conclusions of the original papers – however they report their results.
[6] One might wonder about apparent evidence of discrimination against British immigrants in Canada; indeed, we will see in implementing the correction for the Heckman critique below that this evidence appears to be spurious.

every study finds evidence of discrimination in labor or housing market on the basis of race or ethnicity (Rich, 2014; Zschirnt and Ruedin, 2015; Neumark, 2016; Quillian et al., n.d.), as do the smaller number of studies of discrimination based on sexual orientation (Neumark, 2016). The question this paper addresses is whether this near-uniform evidence of discrimination from field experiments is an accurate reflection of discriminatory behavior, supporting a conclusion that discrimination really is this consistent and pervasive, or whether the evidence in at least some of these studies might reflect biases stemming from differences in the variance of unobservables across groups – the problem highlighted by the Heckman critique.

Some of the studies also include female and male applicants, or more broadly test for discrimination along multiple dimensions, including sex and age (Carlsson and Eriksson, 2014). We do not focus, in this paper, on evidence on discrimination based on sex or age. The broader literature focuses far more on race and ethnicity (and more recently on sexual orientation), and – as we have noted – delivers a near-uniform finding of discrimination against minorities. The evidence of sex discrimination is less robust, and tends to point less to discrimination against women, and more to the importance of sex norms for jobs in whether male or female applicants received more call-backs (Neumark, 2016). And recent evidence from a large-scale correspondence study of age discrimination yields ambiguous results for men, but not women (Neumark et al., 2015).

We next provide a brief discussion of the approach used to correct for the bias in estimates of discrimination from the standard field experiment design, and then present our re-examination of the data from the nine studies we have identified that have the requisite data to implement the method in Neumark (2012) to correct the estimates for bias from differences in the variances of unobservables.

4. Addressing the Heckman critique

There are quite a few critiques of AC studies aside from the one we focus on here. Most of them are laid out in Heckman and Siegelman (1993), and discussed further in Neumark (2012) in the context of the framework laid out in this section. Some of the more important critiques – such as the possibility of "experimenter effects," and small differences between applicants that can matter a lot when applicants are matched on so many characteristics – can be addressed by using correspondence studies instead of audit studies, and indeed most recent research uses the correspondence study technique. The Heckman critique is of particular importance because it applies equally well to correspondence studies, even under otherwise ideal conditions such as no *mean* differences in unobservables between groups, but only differences in the *variances* of unobservables. And this critique is salient because nothing in the research design rules out differences in the variances of unobservables, and indeed – as noted earlier – these differences are foundational in models of statistical discrimination.

We first lay out a basic framework for the analysis of data from an audit or correspondence study, and then explain the bias and the correction.[7] Non-experimental regression-based approaches testing for and measuring discrimination use data on the groups in question in a population, introducing regression controls to try to remove the influence of group differences in the population that can affect outcomes (Altonji and Blank, 1999). Correspondence (and audit) studies, in contrast, create an artificial pool of labor market participants among whom there are supposed to be no average differences by group. This is clearly a potentially powerful strategy, because if we have, e.g., a sample of blacks and whites who are identical *on average*, then in a regression of the form

$$Y = \alpha + \beta B + \varepsilon , \tag{1}$$

where Y is the outcome and B is a dummy variable for blacks, ε is uncorrelated with B, so that the

[7] This section draws heavily on Neumark (2012), while avoiding many details that a reader can find in that paper.

OLS estimate $\hat{\beta}$ (or simply the mean difference in Y) provides an estimate of the effect of race discrimination on Y.[8]

Of course, most of the earlier regression studies focus on wages, whereas AC studies focus on hiring. If an employer is free to pay a lower wage to blacks, for example, then in the context of the Becker employer discrimination model, why discriminate in hiring? One common interpretation is that there is an equal wage constraint – perhaps due to a minimum wage, or because anti-discrimination laws are more effective at rooting out wage discrimination than hiring discrimination. Alternatively, in the simple model, employers with stronger discriminatory tastes than the marginal employer will discriminate in hiring. As we make clear below, however, this framework does not only detect taste discrimination à la Becker.

To provide a more formal framework, suppose that productivity depends on two individual characteristics (standing in for a larger set of relevant characteristics), $X' = (X^I, X^{II})$, so that productivity is $P(X')$. X^I is what the firm observes, and X^{II} is unobserved by firms. It is simplest, for now, to think of Y as continuous, such as the wage offered, although in fact in AC studies we should think of it as latent productivity leading to a decision to hire/call-back or not.

Define discrimination as

$$Y(P(X'), B{=}1) \neq Y(P(X'), B{=}0)\ . \tag{2}$$

Assume that $P(.,.)$ is additive, so

$$P(X') = \beta_I X^I + X^{II}, \tag{3}$$

where the coefficient of X^{II} is normalized to one as it is unobservable, and

$$Y(P(X'), B) = P + \gamma B. \tag{4}$$

Discrimination against blacks implies that $\gamma < 0$, so that blacks are paid less than equally

[8] For simplicity, the discussion here is couched solely in terms of blacks and whites.

productive whites.

In correspondence studies, researchers create resumes that standardize the productivity of applicants at some level. Denote expected productivity for blacks and whites, based on what the firm observes, as P_B^* and P_W^*. Y is observed for each tester, so each test – the outcome of applications to a firm by one black and one white tester/applicant – yields an observation

$$Y(P_B^*, B = 1) - \mathrm{Y}(P_W^*, B = 0) = P_B^* + \gamma - P_W^*. \tag{5}$$

Given that the correspondence study design sets $P_B^* = P_W^*$, we should be able to estimate γ easily from these data, by simply running a regression of Y on the dummy variable B and a constant. (Some potential complications are discussed in Neumark, 2012).

A correspondence study can preclude systematic differences between groups in observables and experimenter effects. But there can still be assumed differences in means between groups despite the groups using matched resumes. In equation (5) above $P_B^* = \mathrm{E}(\beta_I X_B^I + X_B^{II}|X_B^I, B = 1)$, and similarly for P_W^*. Assuming randomization, and with $X_B^I = X_W^I = X^I$, this reduces to $\gamma + \mathrm{E}(X_B^{II}|X^I, B = 1) - \mathrm{E}(X_W^{II}|X^I, B = 0)$, implying that we only identify γ if $\mathrm{E}(X_B^{II}|X^I, B = 1) = \mathrm{E}(X_W^{II}|X^I, B = 0)$. Employers may have different expectations about the mean of X^{II} for blacks and whites, conditional on what they observe, which a labor economist would label statistical discrimination. Although economists are interested in distinguishing between statistical and taste discrimination, both are illegal under U.S. law and both also appear to be illegal under European Union law.[9] Moreover, it is challenging to distinguish between the two models. Thus, this issue is

---

[9] As discussed in Neumark (2016), the U.S. Code of Federal Regulations (29, § 1604.2) defines as illegal discrimination "The refusal to hire an individual because of the preferences of coworkers, the employer, clients or customers …" But it also states "The principle of nondiscrimination requires that individuals be considered on the basis of individual capacities and not on the basis of any characteristics generally attributed to the group. There is not as explicit a prohibition of statistical discrimination in the European Union (EU). Article 2 of the EU's Directive 2000/43/EC prohibits both "direct" and "indirect" discrimination, but these appear to line up, respectively, with disparate treatment and disparate impact in

put aside, and the discrimination estimates from the studies considered in this paper interpreted as the sums of taste and statistical discrimination.[10]

That is not to suggest that researchers using AC methods have not tried to distinguish between taste and statistical discrimination. The idea exploited in most studies is that when the applications include a richer set of applicant characteristics, it is less likely that statistical discrimination plays much of a role in group differences in outcomes (e.g., Ewens et al., 2014). Effectively, one tries to eliminate the term $E(X_B^{II}|X^I, B = 1) - E(X_W^{II}|X^I, B = 0)$ from the estimated difference in hiring rates to see how much of the overall difference in hiring rates is accounted for by this difference in expectations, which corresponds to statistical discrimination.[11] Oreopoulos (2011) presents perhaps the most thorough attempt at discerning between hypotheses about discrimination in an AC study. He uses the approach of adding information (e.g., on country of education, to signal English language skills) to see whether estimated hiring gaps fall, as well as examining differences in hiring gaps across occupations where statistical discrimination is more or less likely to be important. In many cases, he does not find evidence consistent with statistical discrimination, despite evidence from a survey of participating employers that they used name, or country of education or experience, as a signal of potential language problems. One could presumably use the method described below for resumes with varying amounts of information to recover unbiased estimates under different information treatments and hence try to gauge the

---

the U.S. context (see http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A32000L0043, viewed December 2, 2015). However, other material suggests that the prohibition on statistical discrimination is covered by direct discrimination (OECD, 2013, p. 195).

[10] Indeed, it seems that we could also include implicit discrimination (e.g., Bertrand et al., 2005). Implicit discrimination posits a different reason for undervaluing the productivity of a group of workers, which can lead to different policy levers to combat it. But if it arises when employers evaluate applicants in AC studies, the empirical implication for the framework developed here would likely be the same as the implication of taste discrimination.

[11] Neumark (2016) provides many examples, and also some criticisms of this approach.

relative importance of taste and statistical discrimination. However, this issue is not the focus of our analysis in this paper.

The issue raised by the Heckman critique arises from the potential for differences across groups in the variances of the unobservable – which is equally problematic even in the ideal condition of no assumed mean difference. To see how the difference in variances can drive differences in the results of the analysis of data from an AC study, it is most natural to think of equation (1) as a latent variable model for productivity, with applicants having to exceed some productivity threshold with sufficiently high probability (where α in equation (1) can also include observables that vary across individuals that affect productivity, which we have denoted X^I).

To isolate the problem, consider the best-case scenario where $E(X_B^{II}|X^I, B = 1) = E(X_W^{II}|X^I, B = 0)$ – i.e., there is no statistical discrimination regarding levels. But the standard deviations of the unobservables, denoted σ_B^{II} and σ_W^{II}, need not be equal.[12]

Assume the applicant is called back (hired) if there is a sufficiently high probability that their productivity exceeds a given threshold. In this case, the inequality $\sigma_B^{II} \neq \sigma_W^{II}$ combined with the design of AC studies results in a biased estimate of discrimination; worse, we cannot necessarily even sign the bias.

To see the intuition, recall the key feature of the usual design of AC studies of using similar resumes on the applicants in different groups. This requires choosing a particular level of the quality of the resumes. Suppose, for example, that the research design standardizes X^I at a low level, denoted X^{I*}. Employers care about how likely it is that the sum $\beta_I X^I + X^{II}$ exceeds some threshold. Given the low value X^{I*}, this is more likely for a group with a high variance of X^{II}. Thus, even in the case of no discrimination ($\gamma = 0$), the employer will favor the high-variance

[12] Neumark assumes homoscedasticity within groups, and thus suppresses conditioning on X_B^I and X_W^I.

group. Conversely, if standardization is at a high level of X^{I*}, the employer will favor the low-variance group. Because researchers do not have information on the population of real applicants to the jobs studied, there is no definitive way to know whether X^{I*} is high or low relative to the actual distribution, and hence no way to sign the bias.

Neumark (2012) proposed a solution to this problem that separately identifies the relative variances in the unobservables and the discrimination coefficient, γ.[13] The intuition behind the solution stems from the fact that a higher variance for one group (say, whites) implies a smaller effect of observed characteristics on the probability that a white applicant meets the standard for hiring. Thus, information from a correspondence study on how variation in observable qualifications is related to employment outcomes can be informative about the relative variance of the unobservables, and this, in turn, can identify the effect of discrimination. Based on this idea, the identification problem identified by the Heckman critique is solved by invoking an identifying assumption – specifically, that the effect of applicant characteristics that affect perceived productivity and hence call-backs have equal effects across groups – along with the testable requirement that some applicant characteristics affect the call-back probability (since if all the effects are zero we cannot learn about $\sigma_B^{II}/\sigma_W^{II}$ from these coefficient estimates).

In a probit specification, for example, we know that we can only identify the coefficients of the latent variable model for productivity relative to the standard deviation of the unobservable. In this case, we effectively have two probit models, one for blacks and one for whites. If we normalize σ_W^{II} to one, then for a characteristic (Z) that affects the call-back rate, we identify its coefficient (δ_W) relative to σ_W^{II}, or δ_W/σ_W^{II}. However, if we assume that $\delta_W = \delta_B$, then we do not

---

[13] To reiterate, for the purposes of simplification, it is assumed $E(X_B^{II}|X^I, B = 1) = E(X_W^{II}|X^I, B = 0)$. Without this assumption, references to γ in the remainder of this section should be read as references to $\gamma + E(X_B^{II}|X^I, B = 1) = E(X_W^{II}|X^I, B = 0)$ – i.e., the sum of taste and statistical discrimination.

need to impose the normalization that $\sigma_B^{II} = 1$, but instead can identify $\sigma_B^{II}/\sigma_W^{II}$ from the ratio of the coefficients on Z in the probit for whites versus blacks, which in turn allows us to identify γ. The estimation can be done using a heteroscedastic probit model. Finally, when there are *multiple* productivity-related characteristics that shift the call-back probability Z_k (k =1,…, K), there is an overidentification test because the ratio of coefficients on each Z, for whites relative to blacks should equal $\sigma_B^{II}/\sigma_W^{II}$.

The heteroscedastic probit model estimates can be decomposed into the estimated differential due to differences in γ, and the estimated differential due to differences in the variance of the unobservables. In generic notation, let the latent variable depend on a vector of variables S and coefficients ψ, and the variance depend on a vector of variables T, which includes S, with coefficients θ. The elements of S are indexed by k. For a standard probit model, coefficient estimates are translated into estimates of the marginal effects of a continuous variable S variable using

$$\partial P(\text{call-back})/\partial S_k = \psi_k \phi(S\psi) \tag{6}$$

where S_k is the variable of interest with coefficient ψ_k, $\phi(.)$ is the standard normal density, and the standard deviation of the unobservable is normalized to one. Typically, this is evaluated at the means of S. When S_k is a dummy variable such as race, the difference in the cumulative normal distribution functions is often used instead, although the difference is usually trivial.

The marginal effect is more complicated in the case of the heteroscedastic probit model, because if the variances of the unobservable differ by race, then when race "changes" both the variance and the level of the latent variable that determines hiring can shift. As long as we use the continuous version of the partial derivative to compute marginal effects from the heteroscedastic probit model, there is a unique decomposition of the effect of a change in a variable S_k that also appears in T into these two components. In particular, denoting the variance of the unobservable

$[exp(T\theta)]^2$, with the variables in T arranged such that the k^{th} element of T is S_k, then the overall partial derivative of P(call-back) with respect to S_k is

$$\partial P/\partial S_k = \phi(S\psi/exp(T\theta)) \cdot \{\psi_k/exp(T\theta)\} + \phi(S\psi/exp(T\theta)) \cdot \{(-S\psi \cdot \theta_k)/exp(T\theta)\}.^{14} \qquad (7)$$

The first part of the sum in equation (6) is the partial derivative with respect to changes in S_k affecting only the level of the latent variable – corresponding to the counterfactual of S_k changing the valuation of the worker without changing the variance of the unobservable. The second part is the partial derivative with respect to changes via the variance of the unobservable. In the analysis below, these two separate effects are reported as well as the overall marginal effect, and standard errors are calculated using the delta method.[15]

This discussion raises the issue of what we are trying to measure in audit and correspondence studies. Focusing on γ, the structural effect of race, captures the potential discounting by employers of black workers' productivity à la Becker (and possibly statistical discrimination about the mean of X^{II}). But as shown, employers could treat blacks and whites differently in hiring because of different variances of the unobservable. If the latter is accepted as a meaningful measure of discrimination, we might not want to eliminate it.

There are two reasons why the coefficient γ is the focus of interest. First, to the best of our knowledge, differential treatment based on assumptions (true or not) about variances have not been viewed as discriminatory in the legal literature. Second, and probably more important, the taste discrimination (and possibly "first-moment" statistical discrimination) that correspondence

---

[14] See Cornelißen (2005).

[15] Because the formula for the derivative based on a continuous variable yields this unique decomposition, it is used below – and also to interpret the simple probit estimates, as in Table 1. The implied partial derivatives from the probit using the formula for a discrete variable (or computing the partial derivative for each sample observation and averaging, as is now more standard), were very similar. One can decompose the partial derivative from the heteroskedastic probit model based on the partial derivative for discrete variables calculated from difference in the cumulative normal distribution functions, but then the decomposition is not unique.

studies capture in γ generalizes from the correspondence study to the real economy. In contrast, the difference in treatment based on differences in the variance of unobservables is an artifact of the design of correspondence (or audit) studies – in particular, the standardization of applicants to particular, and similar, values of the observables, relative to the actual distribution of observables among real applicants. If, instead, a study used applicants that replicated the actual distribution of applicants to the employers in the study, there would be no bias – in the setting described here – from the different variances of the unobservable; that is, differential treatment is an artifact of the study design. That is not to say, however, that there cannot be discrimination based on second moments with, for example, risk averse firms.

It is rare that correspondence studies include variables that shift the call-back probability, because these studies typically create one "type" of applicant for which there is only random variation in characteristics that are not intended to affect outcomes. However, the nine studies discussed in Section 2 have this information – as in Bertrand and Mullainathan (2004), whose data Neumark (2012) used to illustrate this approach.

5. Results from re-examination of field experiments with quality variation across resumes

Labor market field experiments

We report the results for the re-analysis of the dataset from the labor market field experiments in Tables 2A and 2B. Turning to the first four labor market studies covered in Table 2A, we first report the estimated discrimination coefficient (γ, in the equations from above) in the first row of the table (Panel A). These match the estimates in the last column of Table 1, and have already been summarized.

Panel B turns to the heteroscedastic probit estimates that correct for biases from differences in the variance of unobservables. The "Controls" entry toward the bottom of the table lists the resume characteristics including those likely to shift the call-back rate (like education, skills, etc.).[16] The first row of Panel B reports the overall effect from the heteroscedastic probit estimates. These are similar to the probit estimates. The next two rows of the table report the key results from the decomposition of the heteroscedastic probit estimates. The "level" effect (labelled "Marginal effect through level (unbiased)" in the table) is the unbiased estimate, and the "variance" effect is the artifact of the correspondence study design – which is sensitive to the quality of the resumes sent out relative to the actual distribution, as well as differences in the variances of unobservables.

Looking at these estimates, for the first study – the Baert et al. (2015) experiment on discrimination against Turkish job applicants relative to natives in Belgium – the evidence of discrimination completely disappears in the heteroscedastic probit estimates. In both columns (1) and (2) – the first for a call-back, and the second for an immediate interview – the negative and significant coefficient estimate on the indicator for Turkish applicants becomes positive and statistically insignificant.

In contrast, the estimated effect through the variance is negative and significant, implying that the study design generates bias towards finding evidence of discrimination. The next row of the table reports that the ratio of the estimated standard deviations of the unobservable for minority versus non-minority candidates is around 0.5, indicating a lower variance of unobservables for the Turkish applicants. In terms of the model, the reduction in estimated discrimination coupled with a lower variance of unobservables implies that on average the resumes in this study were of relatively low quality compared to what employers see; thus, the low

[16] Some studies include resume characteristics that are not independent of minority group status. For example, Oreopoulos (2011) indicates, for some of his ethnic groups, that some education or experience occurred in a foreign country. This is useful for asking what might explain variation in the amount of discrimination immigrants face, which is the focus of his study. But it does not fit into the narrower question considered in this paper of discrimination against the minority group per se. Hence, we only use resume characteristics that are constructed to be orthogonal to minority group status.

variance group is less likely to be of sufficiently high quality on the unobservables to merit a call-back, and the difference in variance creates a bias towards finding discrimination against Turkish applicants.

Below the decomposition estimates, the table reports some additional diagnostic test results. First, it reports the p-value from the overidentification test that the ratios of the skill coefficients between (in this case) Turkish and native applicants are equal across all of the skills. The p-value is 0.97 in column (1) and 0.93 in column (2), indicating that we do not reject the overidentifying restrictions. On the other hand, in this case, as reported in the next row, the data tend to reject the restriction to the homoscedastic specification; the p-value from a likelihood ratio test is 0.01 in column (1) and 0.10 in column (2). And the final test result reported is whether the ratio of variances of the unobservables equals one; this is rejected strongly in both columns (a result we expect would to parallel to some extent the likelihood ratio test).

Thus, for the Baert et al. study, application of this method of correcting for bias from differences in the variance of unobservables very much overturns the evidence of ethnic discrimination. There is one additional point to make. One might refer to the negative (and significant) estimates on “Marginal effect through variance” as suggesting that the evidence of discrimination has not gone away, but simply been “displaced” to show up in the variance. We have already explained why, in the context of the method and underlying model used in this paper, the estimated effect through the variance is an artifact of the study, and would not be expected to be replicated in the real world. Similarly, it would not be replicated if the study had used high-quality resumes, or a distribution of resumes that matched the distribution employers actually see. An alternative hypothesis, though, is that the effect of variance is real, and reflects employer risk aversion rather than how the employer evaluates the likelihood that an applicant exceeds a call-back/hiring threshold, given the resume. However, if there is risk aversion, then high-variance groups would be penalized. That is inconsistent with the evidence from the Baert et al. data, since the minority applicants are estimated to have lower variance.[17]

Having gone through the results for the first study in detail, the results for the other labor market studies can be covered more succinctly. The Carlsson and Rooth (2007) study of discrimination against Middle Easterners in Sweden asks a very similar question to Baert et al. (2015). In this case, however, the conclusions are scarcely affected by addressing the Heckman critique. The estimated marginal effect through the level (−0.102) is very similar to the simple heteroscedastic probit estimate (−0.095), and the estimated marginal effect through the variance is close to zero (0.007) and estimated precisely. In this case the ratio of the estimated standard deviations of the unobservable for minorities relative to non-minorities is very close to one (1.03), which implies – in terms of the Heckman critique – that there is unlikely to be any bias regardless of the quality of the artificial resumes relative to the population of resumes that the employer sees, which is certainly consistent with the robustness of the evidence for this study. Note also that the data do not reject the overidentifying restrictions, nor do they reject the restriction to the homoscedastic model or that the ratio of standard deviations equals one – not surprising given the estimates.

The Drydakis (2014) study looks at discrimination against gays and lesbians. In this study, also, correcting for potential bias from differences in the variances of the unobservables does not alter the conclusion much. Indeed, the estimated effect of being gay or lesbian is larger negative (−0.476 or −0.499) after correcting for this bias, relative to the overall effect of −0.384 for gays and −.304 for lesbians. For both groups, the estimated variance of the unobservable is quite a bit

[17] This may be too strong a statement, since if employers actually evaluate applicants based on their assumed variance of the unobservable, the statistical model might be different. We are not aware of any field experiments that have tried to incorporate risk aversion, although this might be fruitful. Dickinson and Oaxaca (2009) provide a lab experiment study of this type of discrimination in labor markets.

larger than for straight men or women, with a ratio of standard deviations of 1.59 for gay versus straight men, and 2.27 for lesbian versus straight women. The combination of a higher variance for gays or lesbians with a larger estimate of discrimination would imply that the resumes were of low quality relative to the distribution, which would lead employers to favor the high variance group and generate a bias towards zero in the estimate of discrimination.

Note that for the Drydakis analyses there is strong evidence against the homoscedastic probit model and marginally significant evidence against equal standard deviations. Also, for the analysis of gay versus straight men the overidentifying restrictions are rejected at the 10-percent level. This last result prompted us to estimate a less restrictive model that did not restrict the effects of two of the resume characteristics to be the same across gay and straight men – chosen based on the estimates indicating that these interactions did not fit the expected pattern if the coefficients in the latent variable model were equal and only the variances of the unobservables varied.[18] In this case the overidentification restrictions were no longer rejected (the p-value was 0.751), yet the estimates were very similar to those reported in column (5) of Table 2A.

Turning to the remaining labor market studies, in Table 2B, Oreopoulos (2011) studies outcomes for six immigrant groups relative to native Canadians. It turns out that for two of these groups – Chinese and Indian – the evidence of discrimination remains significant after addressing the Heckman critique, and is actually stronger, with estimates changing from around −0.05 to −0.10 or greater. For both groups, the estimated variance of the unobservable is larger for

---

[18] These were the indicators for a high-quality resume (more experience) and for resume type. These were chosen because the estimated signs of the interactions relative to the signs of the main effects were rather strongly inconsistent with what would be predicted based on the higher estimated variance of the unobservable for gays. Note that the model is identified as long as the effects of *some* variables that shift the call-back probability are restricted to be equal across the two groups; this restriction does not have to hold for all of them, and can be relaxed by adding interactions between the group indicator and the resume characteristic to the heteroscedastic probit model.

immigrants than for natives, which appears to interact with the applicants being low quality so that the higher variance biases the estimate of discrimination from the standard probit towards zero. In contrast, for the other four groups – Chinese-Canadian,[19] Pakistani, Greek, and British – there is no longer significant evidence of discrimination. Note that in two cases – Pakistani and Greek – the point estimate of the marginal effect of minority group membership through the level is still a large negative number, but is insignificant. In contrast, for the British, the point estimate is no longer negative.

Turning to the other diagnostics, in every case for the Oreopoulos analysis, the overidentification restrictions are not rejected. Similarly, with the exceptions of the analysis for the Chinese applicants, the data do not reject the restriction to the homoscedastic model. Thus, in this case we are sometimes failing to find evidence of discrimination because we are estimating a more flexible model even when the data do not reject a more restrictive model that provides evidence of discrimination – and the results for the Pakistani and Greek applicants are notable in this regard. This poses the usual trade-off of bias versus precision, although generally speaking labor economists are willing to estimate less restrictive models that eliminate bias at the risk of decreased precision. Regardless, it seems reasonable to conclude that the re-analysis of the Oreopoulos data indicates far less robust evidence of discrimination than the original study.

Finally, column (7) of Table 2B repeats the re-analysis of the Bertrand and Mullainathan (2004) data from Neumark (2012). In this case, the evidence of discrimination gets a bit stronger, and the variance of the unobservable is estimated to be larger for blacks. These findings are consistent with low quality resumes generating a bias against finding discrimination, although the qualitative conclusions are unchanged.

[19] This refers to an English first name and a Chinese last name.

Thus, the conclusion from our re-examination of the labor market studies is that the findings from the existing studies of discrimination against ethnic, racial, or sexual orientation minorities are not always robust to addressing the Heckman critique. All 12 estimates based on the existing studies, using the conventional approach, point to evidence of discrimination. But only six (or one-half) of the corrected estimates provide evidence of discrimination.[20]

This conclusion that the analysis of data from field experiments on labor market discrimination is not always robust is echoed in the findings reported in Neumark et al. (2015). They study age discrimination in hiring, and find that the evidence of discrimination against older women is robust to addressing the Heckman critique, but the evidence of discrimination against older men is not robust. On the other hand, some other recent papers using this technique do not find large differences. Carlsson et al. (2013) re-examine data from four previous studies of the Swedish labor market, each of which includes some form of the data required to implement the bias correction. Their re-analysis does not lead to large changes in the estimates of discrimination, although sometimes the estimated discrimination (against those with Arabic names, and in favor of women) becomes smaller. Baert (2015) implemented this method in a study of discrimination against Turkish school-leavers in Belgium, using information on distance from the worker's residence to the workplace and other application characteristics to identify the heteroscedastic probit model, and report that this correction does not alter the conclusions. Thus, among these latter studies, none indicate that ignoring the Heckman critique leads to strong overstatement of discrimination.

Housing market field experiments

[20] This includes the evidence from Carlsson and Rooth (2007), Drydakis (2014, for both gays and lesbians), Oreopoulos (2011, for Chinese and Indian), and Bertrand and Mullainathan (2004, significant at 10-percent level).

The results from the re-examination of the evidence from the housing discrimination studies are presented in Table 3. Ahmed et al. (2010) study discrimination against Arab applicants in Sweden, looking – as three of the four housing studies do – at both positive responses and offers of immediate showings. In this study, correcting for potential bias from differences in the variances of the unobservables does very little to change the conclusions. The estimates of lower positive responses or offer of immediate showing to Arab applicants become if anything more negative – most notably for immediate showing, where the estimate changes from −0.074 to −0.146 – and both estimates are statistically significant. The estimated effects of Arab ethnicity through the variance are positive, and larger for immediate showings, corresponding to the larger negative estimate on the marginal effect through the level. The estimated variance of the unobservable is larger for Arab applicants, so combined, the estimates imply that the applications were lower quality than the population of applications to these landlords, biasing towards zero the conventional probit estimate of discrimination in immediate showings. Turning to the other diagnostics, in neither analysis are the overidentification restrictions, the restriction to a homoscedastic probit model, or equality of the standard deviations rejected. Thus, in this study evidence of discrimination persists.

These same conclusions are echoed in the remaining columns of the table – for the Bosch et al. (2010), Carlsson and Eriksson (2007), and Ewens et al. (2014) studies. In all cases, the bias-corrected estimates still lead to statistically significant evidence of discrimination based on race and ethnicity. And in most cases the point estimate for the marginal effect through the level is very close to the overall heteroscedastic probit estimate, while the estimates of the effect of race or ethnicity through the variance are very small.[21]

There is one case (Ewens et al., 2014) where the overidentifying restrictions are rejected at the 10-percent level (and the p-values for the other tests are fairly low). We therefore carried out an additional analysis, paralleling what we did with the Drydakis (2014) data on gay and straight male applicants. In this case, we estimated a less restrictive model that did not restrict the effects of percent black in the area or city to be the same across black and white applicants, based on the estimates indicating that these interactions did not fit the pattern of equal coefficients in the latent variable model with probit coefficient differing because of differences in the variances of unobservables. In this case the overidentification restrictions were no longer rejected (the p-value was 0.877), yet the conclusions were similar to those in column (7) of Table 3. The overall estimate (standard error) of discrimination from the heteroscedastic probit model was −0.064 (0.023), and the unbiased estimated effect through the level was −0.067 (0.023).

Thus, the conclusion from our re-examination of the housing market studies is that the findings from the existing studies of discrimination against ethnic or racial minorities are robust to addressing the Heckman critique. With one minor exception, these past studies found evidence of discrimination, and our corrected estimates are qualitatively and usually quantitatively very similar.

6. Conclusions

The goal of this paper was to re-examine evidence from field experiments on labor market

---

[21] One reason for the robustness of the results in Carlsson and Eriksson (2013) could be because they use applications with substantial variation in applicant characteristics. The authors do this because by avoiding standardizing applicants to a very narrow range, the bias identified by the Heckman critique can be reduced, although this cannot ensure that the range of quality of actual applicants is not larger. It is also the case that – especially for the positive response outcome – the variances are nearly equal (the ratio of estimated standard deviations is 1.02), so that using a narrow range of applicant quality would not introduce bias.

and housing market discrimination, to see if the near-uniform findings of discrimination against minorities hold up after correcting for an important source of bias originally identified in Heckman and Siegelman (1993) – which we refer to as the "Heckman critique." This critique emphasis that even under quite ideal conditions for these studies, the evidence can be biased in either direction – or, equivalently, discrimination can be unidentified – if the variances of the unobservables differ across the groups studied. This is a plausible concern, given that a difference in the variances of unobservables across cannot be cannot be ruled out and indeed is at the core of early models of statistical discrimination (Aigner and Cain, 1977). We re-examine evidence from nine studies that have the requisite data – applicant or other characteristics aside from the identifier for the group in question which shift the probability of call-backs or hires – implementing a correction for this bias proposed in Neumark (2012).

We find that for the housing market studies, the estimated effect of discrimination is robust to this correction. For the labor market studies, in contrast, the evidence is less robust; in about half of cases the estimated effect of discrimination either falls to near zero or becomes statistically insignificant.

We of course cannot definitively extrapolate from the nine studies we were able to re-examine to the broader set of field experiments on discrimination by race, ethnicity, and sexual orientation. However, given that about half of the estimates of labor market discrimination that we could re-examine no longer provide statistical evidence of discrimination after correcting for bias from differences in the variance of unobservables, it seems reasonable to suggest that the overall (and overwhelming) evidence of labor market discrimination from field experiments is likely less robust than it seems. We have no doubt that in many countries there is discrimination in labor and housing markets against many groups, and that – like the subset of studies we re-examine in this paper – the evidence of discrimination would frequently be robust to addressing the Heckman critique. But our evidence also indicates that in some cases a research design that enables a researcher to address this critique would not find evidence of labor market discrimination. If nothing else, this conclusion implies that we need three types of research to draw more definitive conclusions from field experiments on labor and housing market discrimination: (1) more evidence using this kind of research design and methods; (2) more analysis of how best to implement these methods, what kinds of quality shifters provide the most informative estimates, etc.; and (3) further consideration of whether there are other ways to address the Heckman critique and whether they generate similar answers.

References

Ahmed, A., Andersson, L. and Hammarstedt, M. (2010). 'Can discrimination in the housing market be reduced by increasing the information about the applicants?', *Land Economics*, 86(1), pp. 79-90.

Aigner, D. and Cain, G. (1977). 'Statistical theories of discrimination in labor markets', *Industrial and Labor Relations Review*, 30(2), pp. 175-87.

Altonji, J. and Blank, R. (1999). 'Race and gender in the labor market'. In *Handbook of Labor Economics, Volume 3*, edited by Orley C. Ashenfelter and David Card, 3143-259. Amsterdam: Elsevier.

Baert, S. (2015). 'Field experimental evidence on gender discrimination in hiring: Biased as Heckman and Siegelman predicted?', *Economics*, 9, August 20

http://dx.doi.org/10.5018/economics-ejournal.ja.2015-25.

Baert, S., Cockx, B., Gheyle, N. and Vandamme, C. (2015). 'Is there less discrimination in occupations where recruitment is difficult?', *Industrial and Labor Relations Review*, 68(3), pp. 467-500.

Bertrand, M., Chugh, D., and Mullainathan, S. (2005). 'Implicit discrimination', *American Economic Review Papers and Proceedings*, 95(2), pp. 94-8.

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

Bosch, M., Carnero, M. and Farré, L. (2010). 'Information and discrimination in the rental housing market: Evidence from a field experiment', *Regional Science and Urban Economics*, 40(1), pp. 11-19.

Carlsson, M. and Eriksson, S. (2014). 'Discrimination in the rental market for apartments', *Journal of Housing Economics*, 23, pp. 41-54.

Carlsson, M., Fumarco, L. and Rooth, D.-O. (2013). 'Artifactual evidence of discrimination in correspondence studies? A replication of the Neumark method', IZA Discussion Paper No. 7619. Bonn, Germany: IZA.

Carlsson, M. and Rooth, D.-O. (2007). 'Evidence of ethnic discrimination in the Swedish labor market using experimental data', *Labor Economics*, 14(4), pp. 716-29.

Cornelißen, T. (2005). 'Standard errors of marginal effects in the heteroskedastic probit model', Institute of Quantitative Economic Research, Discussion Paper No. 230. Hanover, Germany: University of Hanover.

Dickinson, D. and Oaxaca, R. (2009). 'Statistical discrimination in labor markets: An experimental analysis', *Southern Economic Journal*, 71(1), pp. 16-31.

Drydakis, N. (2014). 'Sexual orientation discrimination in the Cypriot labor market: Distastes or uncertainty?' *International Journal of Manpower*, 35(5), pp. 720-44.

Ewens, M., Tomlin, B. and Wang, L.-C. (2014). 'Statistical discrimination or prejudice? A large sample field experiment', *Review of Economics and Statistics*, 96(1), pp. 119-34.

Heckman, J. (1998). 'Detecting discrimination', *Journal of Economic Perspectives*, 12(2), pp. 101-16.

Heckman, J. and Siegelman, P. (1993). 'The Urban Institute audit studies: Their methods and findings'. In *Clear and Convincing Evidence: Measurement of Discrimination in America*, edited by Michael Fix and Raymond J. Struyk, 187-258. Washington, D.C.:

The Urban Institute Press.

Jowell, R. and Prescott-Clarke, P. (1970). ‘Racial discrimination and white-collar workers in Britain’, *Race*, 11(4), pp. 397-417.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. New York: Columbia University Press.

Neumark, D. (2016). ‘Experimental research on labor market discrimination.’ NBER Working Paper No. 21262. Cambridge, MA: NBER.

Neumark, D. (2012). ‘Detecting discrimination in audit and correspondence studies’, *Journal of Human Resources*, 47(4), pp. 1128-157.

Neumark, D., Burn, I. and Button, P. (2015). ‘Is it harder for older workers to find jobs? New and improved evidence from a field experiment.’ NBER Working Paper No. 21669. Cambridge, MA: NBER.

OECD. (2013). *International Migration Outlook 2013*. Paris: OECD.

Oreopoulos, P. (2011). ‘Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes’, *American Economic Journal: Economic Policy*, 3(4), pp. 148-71.

Pager, D. (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*, 609(1), pp. 104-33.

Quillian, L., Pager, D., Hexel, O. and Midtboen, A. (n.d.). ‘The persistence of racial discrimination: A meta-analysis of field experiments in hiring since 1972’, unpublished paper.

Riach, P. and Rich. J. (2002). ‘Field experiments of discrimination in the market place’, *The Economic Journal*, 112(483), pp. F480-518.

Rich, J. (2014). ‘What do field experiments of discrimination in markets tell us? A meta analysis of studies conducted since 2000’, IZA Discussion Paper No. 8584. Bonn, Germany: IZA.

Zschirnt, E. and Ruedin, D. (2015). ‘Ethnic discrimination in hiring decisions: A meta-analysis of correspondence tests 1990-2015’, unpublished paper.

Table 1: Experimental Studies of Discrimination in Labor and Housing Markets Re-examined

Study (1)	Country (2)	Years (3)	Minority (4)	Outcome (5)	Majority call-back rate (6)	Estimated differential for minority (7)
A. Labor market field experiments						
Baert et al. (2015)	Belgium	2011-12	Turkish	Call-back	.329	-.082 (.034)
				Immediate interview	.190	-.056 (.026)
Carlsson and Rooth (2007)	Sweden	2005-6	Middle Eastern	Call-back	.269	-.095 (.009)
Drydakis (2014)	Cyprus	2010-11	Gay	Call-back	.554	-.410 (.010)
			Lesbian	Call-back	.523	-.411 (.011
Oreopoulos (2009)	Canada	2008	Chinese	Call-back	.142	-.053 (.007)
			Indian	Call-back	.142	-.056 (.007)
			Chinese-Canadian	Call-back	.142	-.063 (.008)
			Pakistani	Call-back	.142	-.073 (.009)
			Greek	Call-back	.142	-.035 (.017)
			British	Call-back	.142	-.031 (.011)
Bertrand and Mullainathan (2004)	United States	2001-2	Black-sounding names	Call-back	.097	-.030 (.006)
B. Housing market field experiments						
Ahmed et al. (2010)	Sweden	2008	Arab/Muslim	Positive response	.514	-.171 (.033)
				Immediate showing	.254	-.091 (.024)
Bosch et al. (2010)	Spain	2009	Moroccan immigrants	Positive response	.590	-.133 (.014)
				Immediate showing	.541	-.135 (.014)
Carlsson and Eriksson (2014)	Sweden	2010-11	Arab	Positive response	.387	-.130 (.012)
				Immediate showing	.271	-.110 (.011)
Ewens et al. (2014)	United States	2009	Black	Positive response	.503	-.090 (.019)

Note: All studies are correspondence studies. Column (7) reports marginal effect from probit models, our estimates, from following tables. In the Oreopoulos study, "Chinese-Canadian" means there was an English first name.

Table 2A: Estimates for Labor Market Discrimination Studies: Full Specifications

Study	Baert et al. (2015), Belgium		Carlsson and Rooth (2007), Sweden		Drydakis (2014), Cyprus	
Outcome	Call-back	Immed. interview	Call-back		Call-back	
Minority group	Turkish, males		Middle Eastern, males		Gay	Lesbian
	(1)	(2)	(3)	(4)	(5)	(6)
A. Estimates from basic probit						
Minority, marginal effect	-.082	-.056	-.095		-.410	-.411
	(.034)	(.026)	(.009)		(.010)	(.011)
B. Heteroscedastic probit model						
Minority, marginal effect	**-.096**	**-.072**	**-.095**		**-.384**	**-.304**
	(.034)	**(.028)**	**(.009)**		**(.040)**	**(.091)**
Marginal effect through level (unbiased)	**.044**	**.073**	**-.102**		**-.476**	**-.499**
	(.068)	**(.087)**	**(.023)**		**(.029)**	**(.016)**
Marginal effect through variance	-.141	-.145	.007		.093	.195
	(.065)	(.093)	(.026)		(.065)	(.104)
Standard deviation of unobservables, minority/non-minority	**.49**	**.55**	**1.03**		**1.59**	**2.27**
Wald test, overidentification, ratios of coefficients equal (p-value)	.97	.93	.87		.09	.64
LR test: standard vs. heteroscedastic probit (p-value)	.01	.10	.80		.06	.01
Wald test, ratio of standard deviations = 1 (p-value)	.00	.03	.79		.18	.16
Controls (job or applicants)	High education, over-educated, distance, vacancy duration, vacancies/unemployed, unemployment, % foreign, % Turkish, city, multiple jobs, average occupation wage, job quality, intensive/moderate customer contact		Unemployment spells, cultural activities, sport, personality, summer experiences, U.S. high school, high education, multiple employers, occupation		Enhanced resume (more information), high quality (more experience), first applicant, resume type, reference type, tester, occupation	
Clustered (within-pair design)	Yes		Yes		Yes	
N	736	736	5,636		4,846	4,194

Note: In Panel A, the marginal effect is based on the standard formula for a discrete variable, with other variables set at sample means. In Panel B, the continuous approximation for marginal effects is used, with the decomposition in equation (8) immediately below. The standard errors for the two components of the marginal effects are computed using the delta method. The only individual controls for which interactions are not introduced are for other demographic groups.

Table 2B: Estimates for Labor Market Discrimination Studies: Full Specifications

Study	Oreopoulos (2011), Canada						Bertrand and Mullainathan (2004), U.S.
Outcome	Call-back						Call-back
Minority group	Chinese	Indian	Chinese-Canadian	Pakistani	Greek	British	Black-sounding names
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Estimates from basic probit							
Minority, marginal effect	-.053	-.056	-.063	-.073	-.035	-.031	-.030
	(.007)	(.007)	(.008)	(.009)	(.017)	(.011)	(.006)
B. Heteroscedastic probit model							
Minority, marginal effect	**-.046**	**-.050**	**-.068**	**-.083**	**-.066**	**-.038**	**-.026**
	(.009)	**(.008)**	**(.009)**	**(.014)**	**(.073)**	**(.013)**	**(.007)**
Marginal effect through level (unbiased)	**-.131**	**-.101**	**-.029**	**-.076**	**-.169**	**.031**	**-.070**
	(.046)	**(.041)**	**(.054)**	**(.078)**	**(.208)**	**(.045)**	**(.040)**
Marginal effect through variance	.086	.052	-.040	-.007	.102	-.068	.045
	(.052)	(.046)	(.054)	(.070)	(.139)	(.052)	(.043)
Standard deviation of unobservables, minority/non-minority	**1.46**	**1.26**	**.84**	**.97**	**1.54**	**.75**	**1.26**
Wald test, overidentification, ratios of coefficients equal (p-value)	.72	.85	.78	.48	.66	.20	.42
LR test: standard vs. heteroscedastic probit (p-value)	.07	.22	.46	.92	.33	.21	.26
Wald test, ratio of standard deviations = 1 (p-value)	.19	.32	.42	.92	.55	.13	.37
Controls (job or applicants)	Extracurricular activities, top-ranked Bachelor's, Master's, occupation, speaking/social/writing skills required, female						Bachelor's, experience and square, volunteer, military service, email address, gaps in work history, work during school, academic honors, computer and other skills, female; in zip code (% high school dropout, college graduate, black, and white, log median household income)
Clustered (within-pair design)	Yes						Yes
N	5,866	6,373	4,468	3,978	3,388	3,934	4,784

Note: In Panel A, the marginal effect is based on the standard formula for a discrete variable, with other variables set at sample means. In Panel B, the continuous approximation for marginal effects is used, with the decomposition in equation (8) immediately below. The standard errors for the two components of the marginal effects are computed using the delta method. The only individual controls for which interactions are not introduced are for other demographic groups. Some skills are specific to immigrant groups and used to distinguish among immigrants (such as specific language fluencies or where experience obtained), and are not included.

Table 3: Estimates for Housing Discrimination Studies

Study	Ahmed et al. (2010), Sweden		Bosch et al. (2010), Spain		Carlsson and Eriksson (2007), Sweden		Ewens et al. (2014), U.S.
Outcome	Positive response	Immediate showing	Positive response	Immediate showing	Positive response	Immediate showing	Positive response
Minority group	Arab/Muslim		Moroccan immigrants		Arabic/Muslim		Black
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Estimates from basic probit							
Minority, marginal effect	-.171	-.091	-.133	-.135	-.130	-.110	-.090
	(.033)	(.024)	(.014)	(.014)	(.012)	(.011)	(.019)
B. Heteroscedastic probit model							
Minority, marginal effect (unbiased)	**-.165**	**-.074**	**-.136**	**-.136**	**-.131**	**-.113**	**-.089**
	(.034)	**(.027)**	**(.017)**	**(.017)**	**(.013)**	**(.011)**	**(.019)**
Marginal effect through level	**-.182**	**-.146**	**-.136**	**-.135**	**-.134**	**-.074**	**-.092**
	(.035)	**(.049)**	**(.018)**	**(.015)**	**(.026)**	**(.034)**	**(.019)**
Marginal effect through variance	.017	.072	.001	-.001	.004	-.039	.003
	(.019)	(.058)	(.004)	(.014)	(.025)	(.035)	(.003)
Standard deviation of unobservables, minority/non-minority	**1.20**	**1.35**	**.91**	**.98**	**1.02**	**.85**	**1.08**
Wald test, overidentification, ratios of coefficients equal (p-value)	.59	.91	.33	.52	.87	.93	.07
LR test: standard vs. heteroscedastic probit (p-value)	.32	.20	.74	.95	.88	.26	.18
Wald test, ratio of standard deviations = 1 (p-value)	.37	.29	.74	.95	.89	.22	.20
Controls (area or applicants)	Enhanced application, rent, space, rooms, metro, company		Enhanced application, rent, rooms, urban, company, female		Jobs, exercise, nightclub, smoker, references, female, age		Mother's estimated education, positive email, negative email, rent, relative rent, rent in area, one BR, cost, % male, % black in area/city, female, Muslim name
Clustered (within-pair design)	No		Yes		No		No
N	959	959	4,716	4,716	5,827	5,827	13,800

Note: In Panel A, the marginal effect is based on the standard formula for a discrete variable, with other variables set at sample means. In Panel B, the continuous approximation for marginal effects is used, with the decomposition in equation (8) immediately below. The standard errors for the two components of the marginal effects are computed using the delta method. The only individual controls for which interactions are not introduced are for other demographic groups.