DISCUSSION PAPER SERIES

IZA DP No. 7619

Artifactual Evidence of Discrimination in Correspondence Studies? A Replication of the Neumark Method

Magnus Carlsson
Luca Fumarco
Dan-Olof Rooth

September 2013

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Artifactual Evidence of Discrimination in Correspondence Studies? A Replication of the Neumark Method

Magnus Carlsson
Linnaeus University

Luca Fumarco
Linnaeus University

Dan-Olof Rooth
Linnaeus University
and IZA

Discussion Paper No. 7619
September 2013

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 7619
September 2013

ABSTRACT

Artifactual Evidence of Discrimination in Correspondence Studies? A Replication of the Neumark Method

The advocates of correspondence testing (CT) argue that it provide the most clear and convincing evidence of discrimination. The common view is that the standard CT can identify what is typically defined as discrimination in a legal sense – what we label *total* discrimination in the current study –, although it cannot separate between preferences and statistical discrimination. However, Heckman and Siegelman (1993) convincingly show that audit and correspondence studies can obtain biased estimates of *total* discrimination – in any direction – if employers evaluate applications according to some threshold level of productivity. This issue has essentially been ignored in the empirical literature on CT experiments until the appearance of the methodology proposed by Neumark (2012). He shows that with the right data and an identifying assumption, with testable predictions, this method can identify *total* discrimination. In the current paper we use this new method to reexamine a number of already published correspondence studies to investigate if their estimate of *total* discrimination is affected by group differences in variances of unobservable characteristics. We also aim at improving the general understanding of to what extent the standardization level of job applications is an issue in empirical work. We find that the standardization level of the job applications being set by the experimenter appear to be a general issue in correspondence studies which must be taken seriously.

JEL Classification: J71

Keywords: correspondence studies, discrimination

Corresponding author:

Magnus Carlsson
Centre for Labour Market and Discrimination Studies
Linnaeus University
391 82 Kalmar
Sweden
E-mail: Magnus.Carlsson@lnu.se

1. Introduction

Correspondence studies have become an increasingly popular method for measuring discrimination in the labor market. The standard correspondence study sends matched pairs of qualitatively identical applications to employers that have advertised a job opening, the only difference being the name of the applicant, which signals group belonging (see Rich and Riach, 2002, for a survey). The degree of discrimination is quantified by calculating the difference in the number of job invitations to interview between the groups. The advocates of this methodology argue that it provide the most clear and convincing evidence of discrimination. Essentially, the arguments are based on the fact that a carefully designed correspondence study should allow the researcher to circumvent the problem with unobserved individual heterogeneity – a common problem in studies using administrative data. Despite this claim, there are uncertainties regarding interpreting the group difference in callback rates for job interviews solely as arising from discrimination. The main purpose of this paper is to investigate if these alternative explanations are important using already published CT experiments.

A first thing to note is that without making strong assumptions the CT method cannot separately identify the mechanisms that drive discriminatory treatment, that is, whether the difference in callbacks across groups arises from taste based discrimination and/or from statistical discrimination (Heckman, 1998). Not being able to decompose these alternative explanations is certainly a drawback if one wants to decide upon policy measures to prevent discrimination in hiring. On the other hand, one might argue that the most policy relevant issue is to provide proofs of discrimination, and both taste based discrimination and statistical discrimination fall under the legal definition of discriminatory practices.[1] In the reminder of this paper we will refer to *total* discrimination as their common impact on the difference in callback rates across groups, since the inability to separate these mechanisms also apply to our analyses.

The focus of this paper is instead on another type of identification problem, potentially having the consequence of not even being able to identify the level of *total* discrimination, which was first pointed out by Heckman & Siegelman (1993, hereinafter HS). They convincingly show that audit and correspondence studies can obtain biased estimates of *total* discrimination – in any direction – if employers evaluate applications according to some threshold level of productivity.[2] The source of this bias in *total* discrimination originates from the design of the correspondence study, more specifically, from the level of productivity being assigned to applications by the experimenter combined with perceived group differences in the variance of unobserved productivity characteristics. In fact, under such a scenario a standard correspondence study could find discrimination when not existing or find no discrimination when it exists, depending on the standardization level being decided upon for the job applications and which group's variance of unobservables dominates. It is important to note that we do not include a difference in callback rates across groups arising from employer perceptions of group differences in the variance of unobservables in *total* discrimination, since this difference is an artifact of the experimental design.

Despite that the idea that variances of unobservables differ across groups has a long tradition in economics, e.g. Aigner and Cain (1977), this issue has essentially been

[1] See Rooth (2010) and Carlsson & Rooth (2012) for attempts to identify preference based discrimination in CT experiments.

[2] Heckman (1998) also discuss this issue.

ignored in the empirical literature on CT experiments until the appearance of the methodology proposed by Neumark (2012).[3] In short, under some identifying assumptions this method makes it possible to back out the relative variance across groups in a heteroscedastic probit model in order to decompose the difference in group belonging on the callback rate into one part that is due to the level of *total* discrimination mentioned above and one part that is due the standardization level and second moment differences in unobservable productivity characteristics across groups.[4] Neumark applies his method on the CT data used in Bertrand & Mullainathan (2004) and finds suggestive evidence that the estimated degree of *total* discrimination in the original paper is slightly underestimated due to Blacks having a relatively higher variance of unobservables combined with the experiment using a low level of standardization for their job applications. Similarly, Baert et al. (2013) find that their results are only marginally affected when using this correction method.

The main contribution of the current paper is twofold. First, we make use of this new method to reexamine a number of already published correspondence studies to investigate if their estimate of *total* discrimination is affected by group differences in variances of unobservable characteristics. Second, our aim is also to improve upon the general understanding of to what extent the standardization level of job applications is an issue in empirical work and if this varies when job applications are designed to be richer in productivity-related characteristics. For our purposes, we take advantage of data from four correspondence studies conducted in the Swedish labor market between 2005 and

[3] This literature most often suggests that this variance is larger for the minority group.

[4] To be clear, the first moment difference in unobservable characteristics is the difference in means and, hence, is related to the legal definition of discrimination through its relation to statistical discrimination.

2007, which all have the design required for implementing Neumark´s method, that is, include random variation in (some) applicant characteristics.[5] There is also a shift in the construction of these experiments over time developing from standard ones into more advanced experiments including a much richer set of productivity related characteristics.

The most advanced CT experiment, i.e. the data set including the richest set of productivity related characteristics, that we reexamine is found in Carlsson & Rooth (2012) and Rooth (2011) where we find an ethnic gap in the probability of a job interview of around ten percentage points. When we apply Neumark´s method to these data the estimate of *total* discrimination is found to be the same indicating that the experimental design and differences in the variance of unobservables is not an issue in this case. However, when we merge the data from two other experiments including fewer worker attributes, which are found in Carlsson & Rooth (2007), Carlsson (2010) and Rooth (2010), we find that the traditional CT method overestimates the level of *total* discrimination. We interpret this as arising from the experimental design using a low standardization level for the job applications combined with employers perceiving a difference in the variance of unobservable characteristics across groups. Interestingly, when we implement the Neumark methodology on this study, we get a similar degree of *total* discrimination as in the main experiment. In a third analysis we also reexamine Carlsson (2011), which studies gender discrimination in hiring. For this data we find a small gap in the probability of an invitation to a job interview in favor of female applicants using the standard CT method. However, this gap fades away once we apply Neumark´s method. Thus, this new analysis suggests that there is no gender

[5] These experiments are found in various sources: Carlsson & Rooth (2007a), Carlsson & Rooth (2007b), Carlsson (2010), Carlsson (2012), Rooth (2010), Rooth (2011).

discrimination (against male applicants) – at least not for the occupations chosen in this experiment. Finally, when conducting a number of heterogeneity we find that the standardization level of the job applications being set by the experimenter might be a general issue in correspondence studies that must be taken seriously.

The remaining of the paper proceeds as follows. The next section explains the HS critique and Neumark´s methodology in more detail. In Section 3 we implement the Neumark methodology on the data from previous CT studies/experiments, while Section 4 concludes the paper.

2. The Heckman and Siegelman critique and the Neumark solution[6]

The Heckman and Siegelman critique

HS mainly criticize audit studies, and do not explicitly discuss correspondence studies.[7] Audit studies are a slightly different type of experiments, where researchers send real persons or actors to job interviews, compared to correspondence studies, where written applications are sent. HS point out – quite correctly – that audit studies are associated with several methodological problems, but many of these problems, such as experimenter effects, do not apply to correspondence studies. The recent trend towards conducting more correspondence studies and fewer audit studies suggests that researchers have responded to the HS critique. However, one of their critiques applies to correspondence studies as well. In the case employers give a callback for interview only to those job

[6] Much of the content of this section is taken directly from Heckman and Siegelman (1993), Heckman (1998) and Neumark (2012). For a more detailed explanation of the issues involved in this section the reader should turn to those articles.

[7] This applies for Heckman (1998) as well.

applicants who surpass a certain hiring threshold, the level of *total* discrimination could be biased and might partly, or fully, reflect the standardization level of the job applications rather than *total* discrimination.[8]

Following HS, this potential identification problem can be understood with a framework that describes how employers decide whom to invite to a job interview when using a threshold when hiring.[9] Imagine that there is a productivity scale – with low values to the left and high values to the right – and that employers decide whether to invite a job applicant to a job interview based on the likelihood that the applicant´s productivity is above a certain productivity threshold on this scale. Further, suppose that applicants obtain productivity based only on three factors: observed variables, unobserved variables, and a "discount factor", which reflects preference based discrimination. Let the hiring rule be such that an applicant is invited to a job interview if the perceived productivity is above a certain productivity threshold *c*.[10] Suppose the perceived productivity for group $g_i \in \{0,1\}$ is given by

$$\beta_1 X_{g_i}^{I} + X_{g_i}^{II} + \gamma g_i \qquad (1)$$

where $X_{g_i}^{I}$ are observed application characteristics for group *i*, which in a correspondence study typically are identical for the groups, with return equal to β_1, $X_{g_i}^{II}$

[8] Before continuing, a few words should be said about empirically setting the standardization level in an experiment. Information on what standardization level the employers use when hiring is of course unknown to the researcher and hence, the researcher can only assume whether it is high or low relative to the pool of applicants for the jobs applied to.

[9] This critique does not apply if hiring is linear in worker productivity.

[10] Following Heckman and Siegelman (1993) and Heckman (1998), we assume that these factors affect productivity in an additive way.

is unobserved characteristics, not contained in the job application, with a return that has been normalized to one, γ is the (negative) preference discrimination coefficient. The probability that a firm invites an applicant that belongs to group g_i is

$$\Pr\left(c \leq \beta_1 X^{I}_{g_i} + X^{II}_{g_i} - \gamma g_i\right) \qquad (2)$$

The observed variables are deterministic and consequently contribute to the likelihood of passing the productivity threshold in a straightforward way by simply moving an applicant from zero to the right on the productivity scale, closer to – or even beyond – the productivity threshold. A similar logic applies to the preference discrimination factor, although this factor only affects applicants in the discriminated group, who are moved to the left on the productivity scale.

Since unobservable characteristics can be expected to be less influential in the hiring situation when many productivity related characteristics are included, the goal for a CT study is to standardize the worker's productivity on as many observable productivity-related characteristics as possible.[11] Still, taste based discrimination (γ) is only identified if the mean of unobserved perceived productivity ($E(X^{II}_{g_i})$) is identical across groups. To what extent that is the case for a particular CT experiment is untestable and hence, a CT study captures the combined effects of taste based discrimination and statistical discrimination, that is, what the law in many countries recognizes as discrimination. As

[11] Although, the inclusion of productivity-related characteristics is not unrestricted and the job applications being used in the CT experiment have to be similar to the ones being used in the market.

mentioned before, this implies that a CT experiment only identifies *total* discrimination, capturing both mechanisms.

We now turn to the case when the variance of unobservables differs across groups and how this impacts upon the identification of *total* discrimination. Suppose that an employer is confronted with job applications from two different groups, which we label Green applicants and Red applicants, where Green applicants are subject to negative attitudes. The group difference in the probability of an invitation between Red and Green applications is given by

$$\Pr\left(c \leq \beta_1 X^{I}_{\text{Red}} + X^{II}_{\text{Red}}\right) - \Pr\left(c \leq \beta_1 X^{I}_{\text{Green}} + X^{II}_{\text{Green}} + \gamma\right) \tag{3}$$

HS makes a necessary parametric assumption about the distribution of the unobserved variables X^{II}_{Red} and X^{II}_{Green} by assuming that they follow a normal distribution. As will be evident below, only a difference in the variance of unobserved variables across groups is of a concern for the identification of *total* discrimination and we therefore focus on the simplest case where both observed and unobserved group averages of productivity are equal across groups and there is no preference based discrimination ($\gamma = 0$). With these assumptions, and after standardizing, equation (3) becomes

$$\Phi\left(\frac{\beta_1 X^{I} - c}{\sigma^{II}_{\text{Red}}}\right) - \Phi\left(\frac{\beta_1 X^{I} - c}{\sigma^{II}_{\text{Green}}}\right) \tag{4}$$

Unless the variances of the unobservables are equal, this expression is different from zero. Further, not even the sign of the expression can be predetermined without knowledge of the standardization level of the job applications (given by X^I). We illustrate this point graphically. In the first scenario the experimenter sets the standardization level of the job applications quite low relative to the other applicants for the jobs, and the productivity based on the observed variables X^I is located to the left of the threshold *c*, see Figure 1a.

*** Figure 1a here ***

If we assume that Green applicants have a higher variance of unobserved variables than Red applicants, then the former are more likely to reach above the threshold due to the more stretched out distribution of unobservables to the right.

The situation when the standardization level of the job applications is instead set high, and the productivity based on observed variables is located to the right of the threshold *c*, is illustrated in Figure 1b.

*** Figure 1b here ***

If we stick with the assumption that Green applicants have a higher variance of unobserved variables than Red applicants they are now less likely to pass the threshold due to a more stretched out distribution of unobservables to the left.

The discussion above shows that in the case of employers using a hiring threshold and information about the distribution of unobserved variables there can be a difference in callbacks across groups in a CT study although there is no preference based discrimination or statistical discrimination. Importantly, this identification issue is not altered if allowing for differences in the means of unobserved variables and/or taste based discrimination but the probability to be hired is shifted to the left for the discriminated group either counteracting or reinforcing the effect from the difference in variances. Separating these sources from one another is the essence of the Neumark method, which we turn to shortly.

A final issue is why there in this setup is a random component at all in whether a job applicant is invited to a job interview. If all employers make the same probability calculation and simply invite the applicant with the highest likelihood of passing the threshold, it should be deterministic who is invited to a job interview and who is not. Obviously, this is not the pattern we see in reality. However, it is straightforward to incorporate a random component into the employers´ decision making such that the error term follow the distribution of unobservables as given above. One way is to assume that employers have firm specific normally distributed thresholds due to productivity differences.[12]

The Neumark solution

The fundamental problem illustrated above is how to separate between what we label as *total* discrimination, that is, statistical discrimination based on differences in the average of unobservables or taste based discrimination, from the type of artifactual statistical

[12] For other alternatives see Neumark (2012).

discrimination occurring due to employers acting on perceived group differences in the variance of unobservables. Solving this problem ultimately requires that we can estimate the group specific variance. Neumark´s insight is that this can be achieved with the right data, that is, data from a correspondence study that have variation in observed applicant characteristics, and an identifying assumption. Using his method he is able to decompose the marginal effect of group belonging into one part that captures *total* discrimination and one part that is an artifact of the standard of the applications, that is, the effect of group belonging working through the group difference in the variance of unobserved productivity characteristics.

Neumark´s methodology consists of two parts. The first implies obtaining a composite estimate of the group difference in the probability of a job interview, reflecting both the effect through *total* discrimination and the effect through the variance of unobservables. The second part of the methodology implies actually decomposing this composite estimate into its two parts, where the aim is to isolate the part that measures *total* discrimination.

To understand the first step of the methodology, recall that (with preference based discrimination against Green applicants) the difference in the probability of a job interview is

$$\Phi\left(\frac{\beta_1 X^{I}_{\text{Red}} - c}{\sigma^{II}_{\text{Red}}}\right) - \Phi\left(\frac{\beta_1 X^{I}_{\text{Green}} + \gamma - c}{\sigma^{II}_{\text{Green}}}\right) \qquad (5)$$

Without loss of generality this expression can be normalized by the standard deviation of the unobserved variables for, say, Red applicants. The result is

$$\Phi\left(\beta_1 X^{I}_{\text{Red}} - c\right) - \Phi\left(\frac{\beta_1 X^{I}_{\text{Green}} + \gamma - c}{\sigma^{II}_{\text{Green/Red}}}\right) \qquad (6)$$

Since the coefficients in the standard probit model only are identified relative to the standard deviation, the difference in intercepts between the two groups is not identified. Instead, this difference reflects both the group effect and the relative standard deviation. Neumark´s strategy for identifying the effect from group belonging is to utilize data from a correspondence study that contains variation in observed variables. Initially, observations of, say, Red applicants, are used to estimate β_1^{Red}. Then, based on the observations of Green applicants $\frac{\beta_1^{\text{Green}}}{\sigma^{II}_{\text{Green/Red}}}$ is estimated. In the next step Neumark invokes the identifying assumption that $\beta_1^{\text{Red}} = \beta_1^{\text{Green}}$ to obtain $\beta_1^{\text{Red}} \div \left(\frac{\beta_1^{\text{Green}}}{\sigma^{II}_{\text{Green/Red}}}\right) = \sigma^{II}_{\text{Green/Red}}$. With knowledge of $\sigma^{II}_{\text{Green/Red}}$ it is straightforward to obtain the composite estimate of group belonging on the probability of a job interview. In practice, Neumark uses the the heteroskedastic probit model for estimation, where the error term has the standard assumption of zero expectations and a variance equal to $\left[\exp(\omega g_i)\right]^2$ and where g_i again is a group indicator.

The second step of the methodology implies decomposing the composite estimate of group belonging into one part that is the group effect through *total* discrimination – the policy relevant part – and another part that is the group effect through the variance of unobservables. Calculating the marginal effect of group belonging when using the heteroscedastic model is somewhat complicated since when the group indicator changes both the variance and the level of the latent variable that determines callbacks shift. If group belonging is treated as a continuous variable, the marginal effect of belonging to, say, the Green group is calculated by taking the derivative of the probability of a job interview with respect to group belonging. Again, separating these two effects is important since the former should not be treated as discrimination.

In the case of a heteroskedastic probit, the marginal effect is

$$\frac{\partial \Pr(\text{hired})}{\partial \mathrm{g_i}} = \phi\left(\frac{\tilde{X}^1 \cdot \tilde{\beta}}{\exp\left(\omega * g_i\right)}\right) * \left[\frac{\left(\gamma - (\tilde{X}^1 \cdot \tilde{\beta}) * \omega\right)}{\exp\left(\omega * g_i\right)}\right] \qquad (7)$$

where $\tilde{X}^1 \cdot \tilde{\beta}$ is $X^1 \cdot \beta$, but also includes the group indicator variable and its coefficient. Neumark shows that this expression for the composite group effect can be decomposed into the two parts of interest. The first part is the group effect through *total* discrimination, holding the variance constant, which is given by

$$\phi\left(\frac{\tilde{X}^1 \cdot \tilde{\beta}}{\exp\left(\omega * g_i\right)}\right) * \left[\frac{\gamma}{\exp\left(\omega * g_i\right)}\right] \qquad (8)$$

The second part, the group effect through the variance of unobservables, holding the effects through total discrimination constant, which is given by

$$\frac{\partial \Pr(\text{hired})}{\partial \mathrm{g_i}} = \phi\left(\frac{\tilde{X}^1 \cdot \tilde{\beta}}{\exp\left(\omega * g_i\right)}\right) * \left[\frac{(\tilde{X}^1 \cdot \tilde{\beta}) * \omega}{\exp\left(\omega * g_i\right)}\right] \qquad (9)$$

An important question is how likely the identifying assumption of equal returns to observed characteristics across groups is to hold? One could easily come up with stories why it is violated. For example, ethnic groups may attend different schools of different quality and therefore have different returns to education. But as Neumark points out, the identifying assumption is more likely to hold for well designed correspondence studies where it is possible to control for the most obvious group differences. For example, in a written application the experimenter can easily choose schools that are located in similar neighborhoods. Given these concerns it is of outmost importance that the identifying assumption has testable predictions, which we return to below.

3. Data[13]

To implement the Neumark strategy we take advantage of data from no less than four different CT studies conducted in the Swedish labor market, which all have random variation in applicant characteristics in one way or another. In these experiments we investigate both ethnic and gender discrimination. For our purposes in this paper, we combine these studies into three data sets which we label experiment A, B, and C. Neumark´s method requires that the used observed variables have a significant effect on the probability of an invitation and that the effect is the same across groups. The set of variables that fulfills this requirement vary across the experiments and, thus, for each experiment we will use a different set of observed characteristics. In this respect the

[13] Since these experiments are explained in detail in other published articles this section is limited to only the most relevant information for implementing the Neumark method.

labels A through C could also be viewed as a ranking with experiment A having most characteristics being varied and experiment C the least.

Experiment A

In Experiment A we focus on ethnic discrimination against applicants with Arabic names and the data was gathered in a field experiment conducted between March and November 2007. This field experiment was designed for analyzing a number of research questions related to individual worker productivity and therefore has a large variation in productivity characteristics of the fictitious job applications. In principle, twelve different variables were randomly assigned to each application.[14] However, not all of them were found to have an effect on the probability of a job interview or to have the same return across groups. In the end we included five variables that fulfilled these requirements in the analysis of the variance of unobservables. The first two variables regard the personality of the candidate, that is, basically following the Big Five taxonomy using the two of its five categories - extroversion and agreeableness, see Borghans et al. (2008). Both variables are coded as dummy variables in the empirical analysis. The third variable captures in what type of neighborhood the applicant lives with a dummy variable that indicates if the applicant lives in an area with a low or high average income. The fourth variable gives the applicant´s previous work experience as the total length being employed (in years), which varies between one and five years. In the empirical analysis this variable is coded with dummies for each year of experience and with one year serving as the benchmark. Finally, the fifth variable measures whether being engaged in

[14] Details of this experiment are found in Carlsson and Rooth (2012) and Rooth (2011).

sport activities or not, with the benchmark not being engaged in sports. Sport activities could further be exercised at two different effort levels: a recreational and a competitive level, and this variable is included as a dummy for each level of sport activity.

During Experiment A all employment advertisements in thirteen selected occupations found on the webpage of the Swedish employment agency were collected. For these advertised jobs, 5,657 applications, 2,837 with a typical native Swedish name and 2,820 with a typical Arabic name, were sent to 3,325 employers. All applications were sent by email; a clear majority of employers posting vacant jobs at this site accept applications by email. Jobs were applied to all over Sweden, but most advertisements were found in the two major cities of Sweden: Stockholm and Gothenburg. Callbacks for interview were received via telephone (voice mailbox) or e-mail.

Experiment B

In Experiment B we consider gender discrimination against female names. Within the same project as Experiment A, it is also possible to analyze gender discrimination, since additionally 2,830 applications with the same design but now with a native female name were sent to employers.[15] Compared to Experiment A we find much fewer individual variables that affect the probability of a job interview and which also have the same return for both men and women. However, there are variables that have a joint effect that fulfill these requirements. To this end we constructed two new combined variables based on the individual variables. We label these new variables *good labor market characteristics* and *good personal characteristics*, and both are simply two indicators. An applicant is defined as have good labor market characteristics if he or she has at least one

[15] Details of this experiment are found in Carlsson (2012).

of the following characteristics: the person has been abroad for one year during high school; the person has at least four years of experience; the person has experience from more than one previous employer; the person has employment at the moment. An applicant with good personal characteristics is defined as an individual that has at least one of the following characteristics: the person is extrovert or the person is agreeable.

Experiment C

In Experiment C we again consider ethnic discrimination against applicants with Arabic names. What we label Experiment C actually consists of observations from two different correspondence studies found in Carlsson & Rooth (2007), Carlsson (2010) and Rooth (2010). What justifies viewing them as a single experiment for our purposes is that both studies have the same design and are conducted roughly during the same time period between 2005 and 2007. In both experiments the job applicants were born in Sweden with either a typical native Swedish or Arabic name that on average were 25-30 years old, had two to four years of work experience in the same occupation as the job applied for and had obtained their education in the same type of school. Also, in both studies the applications consisted of a quite general biography on the first page and a detailed CV of education and work experience on the second page. Finally, a similar routine for receiving responses from the employers were used. Email addresses and a telephone numbers (including an automatic answering service) were registered at a large Internet provider and a phone company for all fictitious applicants.

Despite the similarities between the two studies there is one important factor that distinguishes them. For reasons unrelated to this paper, the applications in the second experiment were calibrated for six of the occupations relative to the characteristics in the first experiment, that is, the quality of the applications in terms of labor market experience and skills were raised in three occupations and lowered in three occupations.[16] These six occupations contain 3,536 observations. This calibration generates the variation in the standards of the applications that we will utilize in the current paper. However, since only one variable was changed in this experiment we are not able to test the identifying assumption of equal returns to characteristics for this experiment.

4. Empirical analysis

As mentioned earlier, Neumark uses the heteroskedastic probit model to implement his method. Based on the estimated coefficients he then computes the composite marginal effect of group belonging, reflecting the sum of (i) the *total* discrimination effect and (ii) the effect via the variance. To this end we follow Neumark's procedure using the heteroskedastic probit model for estimation, and then decompose the estimated composite marginal effect into the two parts of interest and use the delta method to calculate their standard errors.[17]

Neumark also suggests a procedure to test the identifying assumption of equal coefficients of the observed applicant characteristics across groups. To get the intuition behind this test, imagine as before that we estimate the probability of an invitation to a job interview separately for the two groups. Also, the standard deviation for the two

[16] The quality was raised in the following occupations: sales assistants, accountants, and restaurant workers. The quality was lowered in the following occupations: construction workers, motor-vehicle drivers, and business sales assistants.

[17] All estimations are conducted using Stata 12. The code to calculate the marginal effects and their standard errors is available upon request.

groups is normalized such that the standard deviation is equal to unity for one group and the ratio of the standard deviations for the other group, respectively (see Equation 6). Starting with the simplest case, assume that there is only one observed variable that varies in the applications. If the coefficient for the variable is different between the two groups this could either be because the identifying assumption does not hold or because of the relative standard deviation is different from unity, and we cannot distinguish between the explanations. However, with (at least) two observed variables that vary in the applications, it becomes possible to test the null hypothesis of equal coefficients of the observed applicant characteristics. This can be done by first computing the ratios of the two coefficients separately for each group of applicants. For the second group, the relative standard deviation cancels out, since this is a factor in both the dominator and numerator. Therefore, the null hypothesis of equal coefficients can be tested by testing if the two ratios are equal.[18]

Basic results

Table 1 presents some basic results for Experiment A-C. The purpose with this table is to 1) show the estimated degree of discrimination when we do not take into account the potential effect that differences in the variance of unobservables might have, and 2) provide evidence of that the observed application variables that we will rely on when implementing Neumark’s methodology have significant effects – with expected signs – on the probability of an invitation to a job interview.

[18] As Neumark points out, failing to reject the null hypothesis of equal coefficients does not decisively rule out the alternative hypothesis of unequal coefficients. On the other hand, with a large number of varying variables, failing to reject a false null hypothesis becomes less likely.

*** Table 1 here ***

The basic results for Experiment A are in the first two columns of the table. In the first column in the top row we find that the ethnic difference in the probability of a job interview is 9.4 percentage points without control variables. This is the number that would be reported as the main result in a correspondence study capturing the *total* discrimination effect. From the following seven rows of this column it is evident that applicants that are extrovert, agreeable, or have more than one year of experience (the benchmark) have significantly higher probability of receiving an invitation to a job interview. Also, the next two rows in this column show that applicants that are engaged in sport activities have a (weakly significant) higher probability of an invitation to a job interview. This means that essentially all the observed application variables have a significant effect – with the expected signs – on the probability of a job interview. While the regression underlying the estimates in the first column does not include any other control variables, the second column includes other application attributes and occupational fixed effects. The other application controls include dummy indicators for whether the job applied for was located in Stockholm, Gothenburg, or in other parts of Sweden, the order the applications were sent, and the typeface and layout of the application.[19]

The basic results for Experiment B are found in the next two columns of the table. Again, in the top row we find the group difference in the probability of a job interview,

[19] The fact that the reported marginal effects of the observed application variables are more or less identical in the two columns suggests that the randomization of the application variables have succeeded.

now between male and female applicants, which is 2.6 percentage points in favor of female applicants. From the estimates further down in the table it is evident that applicants that have good labor market and personal characteristics have a significantly higher probability of an invitation to a job interview (three and four percentage points, respectively).[20]

Finally, the basic results for Experiment C are found in the last column of Table 1. This time the ethnic difference in the probability of a job interview is 12.8 percentage points, in favor of applicants with native Swedish names. The row at the bottom of the table reveals that improved quality applications have a significantly higher probability of a job interview by four percentage points. Note that there is only a single column with estimates for Experiment C. This is partly because in this experiment we do not have any useable information to construct other application controls other than high quality. Moreover, in the case of Experiment C it does not make sense to present the estimates without occupational fixed effects. The reason is that the quality of the applications where manipulated at the occupational level, which means that without occupational fixed effects the estimate of improved quality will also reflect job specific demand.

Main results

Table 2 presents the baseline results for the Neumark decomposition. The first row is for comparison and repeats the first row of Table 1 and is estimated using the dprobit

[20] The fact that the estimated marginal effects of the observed application characteristics are unaffected in this case too by whether or not other application controls and occupational fixed effects are included in the regressions provides further evidence for that the randomization procedure have succeeded.

command in Stata. The results in the following rows of the table present the main results of the paper.

*** Table 2 here ***

Considering ethnic discrimination in Experiment A (first column), the second row shows the estimated composite marginal effect of having a typical Arabic name being obtained from the heteroskedastic probit model. The fact that this estimate is very similar to the estimate from the standard probit suggests that differences in the variance of unobservables might not be an important issue in this experiment. The next two rows give the marginal effects of group belonging decomposed into the effect through the level (*total* discrimination) and through the variance of unobservables. For Experiment A the marginal effect through the level is very similar to the composite marginal effect from group belonging and hence, there is as expected no evidence of an effect through the variance. Also, the point estimate of the relative standard deviation for applicants with typical native Swedish and Arabic names is very close to unity (.96).[21] This implies that the standard CT methodology showed an unbiased measure of *total* discrimination, that is, the sum of taste based and statistical discrimination.

Interestingly, when considering gender discrimination (Experiment B, column 2) the composite group effect in favor of females is somewhat higher (see second row) than what was found in with the standard probit model (first row). In other words, when using the standard CT methodology for this experiment the level of *total* discrimination is

[21] Further, the high p-value for the Wald statistic on the last row suggests that the data is consistent with the identifying assumption of equal coefficients for the observed applicant characteristics.

estimated as somewhat lower than the estimate for the measure of group belonging using the heteroscedastic probit model. Also, the estimates from the decomposition found in the next two rows indicate that the composite effect goes entirely through the effect of the variance of unobservables. This implies that a standard correspondence study measure of *total* discrimination would find spurious evidence of discrimination against males. Although the relative standard deviation is not different from one in a statistical sense (p-value: .29) the interpretation of the point estimate is that the standard deviation of the unobserved variables is 13 percent higher for females compared to males. This is consistent with a low standard of the applications being set in the experiment and where the higher variance of unobservable charactersitics benefits females. The high p-value for the Wald statistic on the last row suggests that the data is consistent with the identifying assumption of equal coefficients for the observed applicant characteristics.

Finally, for Experiment C (ethnic discrimination, last column) the standard probit model shows a smaller marginal effect of having a typical Arabic name compared to the marginal effect of the same group belonging in the heteroscedastic probit (compare the first and second row). Indeed, the estimates found in the next two rows suggest that quite a large fraction of the composite effect of group belonging go through the effect of the variance of unobservables. This implies that a standard correspondence study estimate of *total* discrimination overestimates the degree of discrimination against applicants with typical Arabic names. Interestingly, the point estimate of the marginal effect of *total* discrimination using the Neumark decomposition (third row) is very similar to what was found for Experiment A. This indicates the importance of actually implementing the Neumark method when comparing the estimate of *total* discrimination across experiments using different designs, since the use of different standardization levels, i.e. using a different set of productive characteristics across experiments, together with employers acting on perceived differences in the variance unobservables would hide similarities in this estimate. In principle, in this case the Neumark method simply adjusts each experiment for the standardization level of the job applications.

Although the relative standard deviation in experiment C is not different from one in a statistical sense (the p-value is .14) the interpretation of the point estimate is that the standard deviation of the unobserved variables for applicants with typical Arabic names is only .83 of the standard deviation for applicants with native Swedish names. Similarly as for Experiment B, this is consistent with setting a low standard of the applications in the experiment where applicants with typical Arabic names suffering from their lower variance of the unobservables.

Heterogeneity

The average standard of the pool of applications might in realty vary across occupations. In other words, what is a relatively low standard in one occupation might be a relatively high standard in another occupation and vice versa. Similarly, it is not difficult to imagine that differences in the variance of unobservables also vary across occupations. If the effect of these varying factors goes in opposite directions for different occupations interesting patterns might be hidden in the analysis of the total sample. While we do not have enough observations to analyze single occupations, it is for Experiment A and B possible to do the analysis on subsamples based on the type of occupation. However, this is not possible for Experiment C where the observed application characteristics where manipulated at the occupational level and there are relatively few occupations. In this section we repeat the analysis in Table 2 for Experiment A and B, but separately for low/high skilled occupations, occupations with a share of immigrants above/below the population average, and a share of females above/below the population average. Also, for Experiment A, where we find the most compelling differences in callbacks, we divide the occupations into three groups of occupations based on the job specific demand.

*** Table 3 here ***

Table 3 presents the results for low/high skilled occupations. A high skilled occupation is defined as a job that requires a university degree.[22] Starting with Experiment A (ethnic discrimination), the results for low skilled jobs are very similar to the results for the whole sample; there is no evidence for that differences in the variance of unobservables is an issue. For high skilled jobs, however, the results indicate that the composite group effect is slightly higher than what was found with the standard probit model. When this effect is decomposed (see the next two rows), we see that the effect partly goes through the effect of the variance of unobservables. Thus, a standard correspondence study would overestimate the degree of ethnic discrimination for high skilled jobs. Again, the relative standard deviation is not different from one in a statistical sense (p-value: .29), but the interpretation of the point estimate is that the standard deviation of the unobserved variables is lower for applicants with typical Arabic names. This is consistent with a low standard of the applications being set in the experiment and

[22] The high skilled occupations are accountants, primary school teachers (math/science), high school teachers, computer professionals, nurses, and primary school teachers (language)

where the lower variance of unobservable characteristics is to the disadvantage of applicants with a typical Arabic name.

For Experiment B (sex discrimination) the results for both low and high skilled occupations are similar to what was found for the total sample. Total discrimination (against males) appears to largely go through the variance of unobservables.

Next, the results for occupations with a relatively large/small share of immigrants/females are reported in Table 4, where the occupations simply are divided around the average share of immigrants and females in the population.[23] In the case of ethnic discrimination (Experiment A), we explore heterogeneity with respect to the share of immigrants, and in the case of sex discrimination (Experiment B), we explore heterogeneity with respect to the share of females. The motivation is that we find it likely that ethnic differences in the variance of unobservables vary to a larger extent with the share of immigrants in an occupation, while sex differences in the variance of unobservables vary to a larger extent with the share of females. The argument is based on the idea that the own ethnic group/gender might have better/different information about the unobservables of an applicant in an occupations.

*** Table 4 here ***

[23] In the total population, the share of females is around 50 % and the share of immigrants was 13.4 % in 2007 (the latter figure is taken from Statistics Sweden, see http://www.scb.se/Pages/ProductTables.aspx?id=25795).

For ethnic discrimination (Experiment A) the results in the first column suggest that the basic probit underestimates the composite effect of ethnic discrimination in occupations with a high share of immigrants compared to the heteroskedastic probit (see row 1 and 2). When the composite effect for the heteroskedastic probit is decomposed, the effect through the level (*total* discrimination) dominates, but it appears that a portion also goes through the variance. The interpretation is that an analysis of these occupations that does not implement Neumark´s methodology would overestimate the degree of ethnic discrimination. For occupations with a share of immigrants below average differences in the variance of unobservables does not appear to be a major issue.

For gender discrimination (Experiment B), the results for occupations with a share of females above average are very similar to what was find in the main analysis; females have a small advantage based on the composite measure, but this effect appear to entirely go through the variance (third column). The results for occupations with a share of females below average are even more interesting (last column). The composite effect based on the heteroskedastic probit suggests that having a female name increases the probability of an invitation with approximately three percentage points. However, decomposing this effect suggests that based on *total* discrimination females in male dominated occupations are actually at a disadvantage. Taking the estimate at face value (about 11 percentage points), discrimination against females in this case is as large as what we typically find for Arabic names. Of course, this estimate must be interpreted with caution, since the precision is very low. But at least the estimate indicates that differences in the variance of unobservables might be problematic in CT. Considering the relative standard deviation, although not statistically significant, the interpretation of the estimate of the relative standard deviation is that the standard deviation of the unobservables is larger by a factor 2.42 for female applicants compared to male applicants in male dominated occupations. Again, this is consistent with a low standard of the applications, where females have an advantage of their higher standard deviation of unobservables.

Finally, we investigate heterogeneity with respect to job specific demand. In this case, we only show the results for experiment A, where we find the largest initial differences in callback rates. As an indicator of job specific demand, we use the occupational callback rate for female applicants found in experiment B. We identify three groups of jobs, containing jobs which are homogenous with respect to the callback rate for female applicants. The first group of occupations with a relatively low demand (<= 22 percent callback rate for female applicants) consists of shop sales assistants, cleaners, and teachers (language). The second group, with occupations with a medium demand (> 22 and <= 32 percent callback rate for female applicants), contains mechanics, construction workers, accountants, high school teachers, business sales assistants, and teachers (science). The last group, with occupations with relatively high demand (> 32 percent callback rate for female applicants), consists of restaurant workers, motor vehicle drivers, computer professionals, and nurses.

*** Table 5 here ***

Table 5 presents the results for the three groups of occupations. Considering the group with a low demand (first column), we see that the composite marginal effect of having an

Arabic name is somewhat higher with the heteroskedastic probit compared to the standard probit. The next two rows, which show the results from the decomposition, suggest that the major part of the composite effect goes through the variance of unobservables. This means that a correspondence study that does not implement Neumark´s method would overestimate *total* discrimination in occupations with a low demand. Similarly as before, Arabic names appear to have a lower (although not significant) variance of unobservables. Again, the results are consistent with a low level of standardization, where Arabic names are at a disadvantage because of their lower variance of unobservables. Next, for occupations with medium demand the results are very similar (second column). In contrast, for occupations with a high demand (last column) the decomposition of the composite estimate of discrimination indicates that a CT that does not take differences in the variance of unobservables into account might slightly underestimate *total* discrimination. In this case, the interpretation of the point estimate of the relative standard deviation is that Arabic names have a higher variance. This is again consistent with a low level of standardization, but in this case to the advantage to applicants with Arabic names through their higher variance of unobservables.

5. Discussion and concluding remark

Many researchers hold the view that correspondence studies provide the most clear and convincing evidence of discrimination, since these studies cleanly take care of any omitted applicant characteristics that might be correlated both with the applicant´s group belonging and employment opportunities. However, correspondence studies might obtain biased estimates of *total* discrimination, in any direction, if employers evaluate applications in a non-linear way by giving a callback to interview to all job applicants that have qualifications above a certain threshold. In this non-linear setting, the standardization level of the applications together with group differences in the variance of unobservables can generate the bias in *total* discrimination. In fact, correspondence studies could indicate discriminatory practices when not existing or find no discrimination when it exists.

However, the empirical literature on CT experiments has until the recent methodology proposed by Neumark (2012) ignored this issue. In the current paper we apply this new method with data from a number of previously published CT studies, which have the necessary ingredients for implementing Neumark´s method. Our first purpose is to reexamine these studies and investigate if the conclusion for the level of *total* discrimination is biased. The second goal is to further improve the understanding of to what extent employers use information about group differences in variances of unobservables and to what extent the Heckman and Siegleman critique is an important issue in empirical work.

The first study that we reexamine contains a large number of individual productivity characteristics that are experimentally varied. We find an ethnic gap in the probability of a job interview of about ten percentage points when employing the standard CT approach. This estimate of *total* discrimination is unaltered when adopting the Neumark method, indicating that perceived group differences in the variance of unobservables are not important for hiring in this experiment. However, the results from a reexamination of a second CT experiments shows a different picture indicating that the HS critique is not groundless. In this experiment, in which a somewhat higher degree of *total* discrimination is found using a standard CT approach, we find this estimate to overestimate the true level. In fact, when we implement the Neumark methodology to purge the *total* discrimination estimate off the effect from the two experiments using different standardization levels, we find *total* discrimination to be the same in the experiments. Moreover, based on a completely unrelated (and unpublished) correspondence study we again find a similar degree of ethnic discrimination as in Carlsson & Rooth (2007), when we apply Neumark´s methodology. A similar result applies when we reinvestigate a CT study on gender discrimination. Using the standard CT approach there is small gender gap in the probability of an invitation to a job interview in favor of female applicants. However, when we apply Neumark´s method this difference appears to be entirely driven by the variance of unobservables. Thus, Neumark´s method suggests that gender discrimination (against males) is overestimated.

Finally, by conducting a number of heterogeneity analyses, we find further support for the standardization level of the job applications indeed being an issue in empirical work with correspondence testing. The most striking result is found when we investigate gender discrimination in male dominated occupations. Although the precision is low, the results indicate that there is no gender gap using a standard CT framework, but females are at a potentially large disadvantage when we take the standardization of the applications and group differences in the variance of unobservables into account.

We conclude, in line with Neumark, that it seems important that future correspondence studies of discrimination incorporate random variation in observed variables to facilitate an analysis of to what extent the results are affected by the standard of the applications.

References

Altonji, Joseph G. & Blank, Rebecca M. (1999), "Race and gender in the labor market," Handbook of Labor Economics, in: O. Ashenfelter & D. Card (ed.), Handbook of Labor Economics, edition 1, volume 3, chapter 48, pages 3143-3259 Elsevier.

Ahmed A. M., Andersson L., Hammarstedt M. (2010), "Can discrimination in the Housing Market Be Reduced by Increasing the Information about the Applicants?", Land Economics, Vol. 86, No. 1, pp. 79-90.

Aigner D. J., Cain G. G. (1997), "Statistical Theories of discrimination in Labor Markets", Industrial and Labor Relations Review, vol. 30, No. 2, pp. 175-187.

Allison P. D. (1999), "Comparing Logit and Probit coefficients Across Groups", Sociological Methods & Research, Vol. 28, No. 2, pp. 186-208.

Altonji J. G., Pierret C. R. (1997), "Employer Learning and Statistical Discrimination", NBER Working Paper 6279, National Bureau of Economic Research (NBER), pp. 1-64.

Antonovics K., Arcidiacono P., Walsh R. (2005), "Games and Discrimination: Lessons From The Weakest Link", Journal of Human Resources, Vol. 40, No. 4, pages 918-947.

Antonovics K., Knight B. G. (2009), "A New Look at Racial Profiling: Evidence from the Boston Police Department", The Review of Economics and Statistics, Vol. 91, No. 1, pages 163-177.

Arrow K. J. (1973), "The Theory of Discrimination", Industrial Relations Section, Princeton University, Working Paper No. 30A, pp. 1-31.

Arrow K. J. (1998), "What has economics to say about racial discrimination", The Journal of Economic Perspectives, Vol. 12, No. 2 , pp. 91-100.

Ashenfelter O., Hannan T. (1986), "Sex Discrimiantion and Product Market Competition: The Case of Banking Industry", The Quarterly Journal of Economics, Vol. 101, No. 1, pp. 149-174.

Ashenfelter O., Oaxaca R. (1987), "The Economics of Discrimination: Economists Enter the Courtroom", The American Economic Review, Vol. 77, No. 2, Papers and Proceedings of the Ninety-Ninth Annual Meeting of the American Economic Association, pp. 321-325.

Ayres I., Siegelman P. (1995), "Race and Gender Discrimination in Bargaining for a New Car", The American Economic Review, Vol. 85, No. 3, pp. 304-321.

Baldini M., Federici M. (2011), "Ethnic discrimination in the Italian rental housing market", The Journal of Housing Economics, Vol. 20, pp. 1-14.

Baert, S., Cockx, B., Gheyle, N. & Vandamme, C. (2013) "Do Employers Discriminate Less If Vacancies Are Difficult to Fill? Evidence from a Field Experiment", IZA Discussion Paper #7145.

Becker G. (1971), “The Economics of Discrimination”, 2nd Ed., Chicago: University of Chicago Press.

Bertrand M., Mullainathan S. (2004), “Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination”, The American Economic Review, Vol. 94, No. 4, pp. 991-1013.

Blinder A. S. (1973), “Wage discrimination: Reduced Form and Structural Estimates”, The Journal of Human Resources, Vol. 8, No. 4, pp 436-455.

Carlsson M., D.-O. Rooth (2007), “Evidence of Ethnic Discrimination in the Swedish Labor Market Using Experimental Data”, Labour Economics, Vol. 14, No. 4, pp. 716-729.

Carlsson M. (2010), “Experimental Evidence of Discrimination in the Hiring of First- and Second-generation Immigrants”, LABOUR, CEIS, Fondazione Giacomo Brodolini and Wiley Blackwell Ltd, Vol. 24, No. 3, pp. 263-278.

Carlsson & Rooth (2012) "Revealing taste-based discrimination in hiring: a correspondence testing experiment with geographic variation," Applied Economics Letters, vol. 19(18), pp1861-1864.

Cornelißen T. (2005), “Standard Errors of Marginal Effects on the Heteroskedastic Probit Model”, University of Hannover, Institute of Quantitative Economic Research, Discussion Paper No. 320, pp. 1-11.

Dickinson D. L., Oaxaca R. L. (2006), “Statistical Discrimination in Labor Markets: An Experimental Analysis”, IZA Discussion Papers 2305, Institute for the Study of Labor (IZA), pp. 1-27.

Ewens M., Tomlin B., Wang C. (2012), “Statistical Discrimination or Prejudice? A Large Sample Field Experiment”, Carnegie Mellon University, Department of Economics, Discussion Paper No. 23/12, pp. 1-56.

Falk A., Fehr E. (2003), “Why Labour Market Experiments?”, Labour Economics, Vol. 10, No. 4 , pp. 399-406.

Fershtman C., Gneezy U. (2001), “Discrimination in a Segmented Society: An Experimental Approach”, The Quarterly Journal of Economics, Vol. 116, No. 1, pp. 351-377.

Goldberg P. K. (1996), “Dealer Rpice Discrimination in New Car Purchases: Evidence from the Consumer Expenditure Survey”, Journal of Political Economy, vol. 104, No. 3, pp. 662-654.

Goldin C., Rouse C. (2000), “Orchestrating Impartiality: The Impact of Blind Auditions on Female Musicians”, American Economic Review, Vol. 90, No. 4, pp. 715-741.

Harvery A. C. (1976), “Estimating Regression Models with Multiplicative Heteroskedasticity”, Econometrica, Vol. 44, No. 3, pp. 461-465.

Heckman J. J., Siegelman P. (1993), “The Urban Institute Audit Studies: Their Methods and Findings”, in: Fix M., Struyk R. (1993), “Clear and Convincing Evidence: Measurement of Discrimination in America”, Washington DC: The Urban Institute Press, pp. 7-258.

Heckman J. J. (1998), “Detecting Discrimination”, Journal of Economic Perspectives, Vol. 12, No. 2, pp. 101-116.

Jann B. (2008), “The Blinder-Oaxaca decomposition for linear regression models”, The Stata Journal, Vol. 8, No. 4, pp. 453-479.

Jowell R., Prescott-Clarke P. (1970), “Racial discrimination and white collar workers in Britain“, Race, Vol. 11, No. 4, pp. 397-417.

Kahn L. M., Sherer P. D. (1988), “Racial Differences in Professional Basketball Players’ Compensation”, Journal of Labor Economics, Vol. 6, No. 1, pp. 40-61.

Klumpp T., Su X. (2012), “Second-Order Statistical Discrimination”, Emory University working paper, pp. 1-23.

Knowles J., Persico N., Todd P. (2001), “Racial Bias in Motor Vehicle Searches: Theory and Evidence”, Journal of Political Economy, Vol. 109, No.11, pp. 203-229.

Levitt S. D. (2003), “Testing Theories of Discrimination: Evidence from Weakest Link”, NBER Working Paper 9449, National Bureau of Economic Research (NBER), pp. 1-22.

List J. A. (2004), “The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field”, The Quarterly Journal of Economics, Vol. 119, No. 1, pp. 49-89.

Maynard D. C., Taylor E. B., Hakel M. D. (2009), “Applicant Overqualification: Perceptions, Predictions, and Policies of Hiring Managers”, in: Chen O. T. (2009), “Organizational Behavior and Dynamics”, Hauppauge, NY: Nova Science Publishers, pp. 13-38.

Neal D. A., Johnson R. W. (1996), “The Role of Premarket Factors in Black-White Wage Differences”, The Journal of Political Economy, Vol. 104, No. 5, pp. 869-895.

Neumark D (2010), “Detecting Discrimination with Audit and Correspondence Studies”, IZA Discussion Papers 5263, Institute for the Study of Labor (IZA), pp. 1-42.

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

Oettinger G. S. (1996), “Statistical Discrimination and the Early Career Evolution of the Black- White Wage Gap”, Journal of Labor Economics, Vol. 14, No. 1, pp. 52-78.

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**,** Vol. 609, No. 1, pp. 104-133.

Phelps E. S. (1972), “The Statistical Theory of Racism and Sexism”, The American Economic Review, Vol. 62, No. 4, pp. 659-661.

Political and Economic Planning, “Report on Racial Discrimination”, London: Political and Economic Planning, 1967.

Riach P. A., Rich J. (1991-2), “Measuring Discrimination by Direct Experimental Methods: Seeking Gunsmoke”, Journal of Post Keynesian Economics, Vol. 14, No. 2, pp. 143-150.

Riach P. A., Rich J. (2002), “Field Experiments of Discrimination in the Market Place”, The Economic Journal, Vol. 112, No. 482, pp. 480-518.

Rooth D.-O. (2010), “Automatic Associations and Discrimination in Hiring: Real World Evidence”, Labour Economics, Vol. 17, No. 3, pp. 523-534.

Sidanius J., Pratto F. (1999), “Social dominance: An intergroup theory of hierarchy and oppression”, New York: Cambridge University Press.

Sidanius P., Veniegas R. C. (2000), “Gender and Race Discrimination: The Interactive Nature of Disadvantage”, in Oskamp S. (2000), “Reducing Prejudice and Discrimination The Claremont Symposium on Applied Social Psychology”, Mahwah, New Jersey: Lawrence Erlbaum Associates, pp 47-69.

Yinger J. (1986), “Measuring Racial Discrimination with Fair Housing Audits: Caught in the Act”, The American Economic Review , Vol. 76, No. 5, pp. 881-893.

Ward R. (1969), “A Note on the Testing of Discrimination”, Race & Class, Vol. 11, No. 2, pp. 218-223.

Williams R. (2009), “Using Heterogeneous Choice Models to Compare Logit and Probit coefficients Across Groups”, Sociological Methods & Research, Vol. 37, No. 4, pp. 531-559.

Figures and tables

Figure 1a. Low level of standardization.

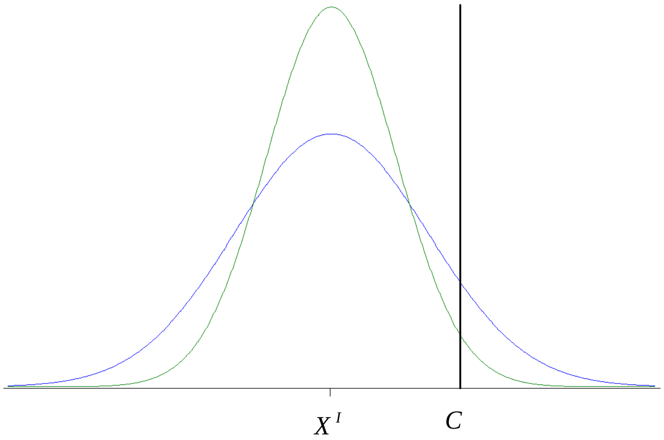

Figure 1b. High level of standardization.

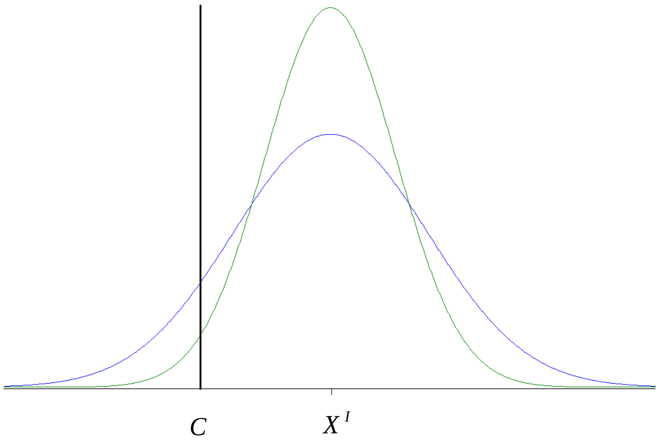

Table 1. Basic probit.

	Ethnicity		Gender		Ethnicity
	Experiment A		Experiment B		Experiment C
	(1)	(2)	(1)	(2)	(2)
Arabic/Female Name	-.094***	-.096***	.026***	.026**	-.128***
	[.009]	[.009]	[.009]	[.009]	[.010]
Application characteristics					
Extroversion/competence	.03***	.04***	-	-	-
	[.01]***	[.01]			
Agreeableness	.03**	.02**	-	-	-
	[.01]	[.01]			
Experience = 2	.02	.03	-	-	-
	[.02]	[.02]			
Experience = 3	.06***	.06***	-	-	-
	[.02]	[.02]			
Experience = 4	.08***	.08***	-	-	-
	[.02]	[.021]			
Experience = 5	.03	.04**	-	-	-
	[.02]	[0.02]			
Bad neighborhood	-.02*	-.02	-	-	-
	[.01]	[.01]			
Recreational sports	.02	.01	-	-	-
	[.01]	[0.01]			
Competitive sports	.03*	.03	-	-	-
	[.02]	[0.02]			
Good labor market characteristics	-	-	.03***	.03***	-
			[.01]	[.01]	
Good personal characteristics	-	-	.04***	.04***	-
			[.01]	[.01]	
Increased quality application	-	-	-	-	.04***
					[.02]
Other application controls	No	Yes	No	Yes	-
Occupational fixed effects	No	Yes	No	Yes	Yes
N	5,636	5,636	5,662	5,662	3,536

Notes:

Table 2. Decomposition.

	Ethnicity Experiment A	Gender Experiment B	Ethnicity Experiment C
A. Basic probit			
Arabic/Female name	-.096***	.026**	-.128***
	[.009]	[.009]	[.010]
B. Heteroskedastic probit			
Arabic/Female name (unbiased)	-.097***	.029***	-.133***
	[.009]	[.010]	[.011]
Marginal effect of name through level	-.088***	.001	-.090**
	[.028]	[.024]	[.036]
Marginal effect of name through variance	-.010	.028	-.044
	[.025]	[.025]	[.033]
Relative standard deviation of unobservables	.96	1.13	.83
Wald test statistic, standard deviation == 1 (p-value)	.68	.29	0.14
Wald statistic, ratios of coefficients are equal (p-value)	.67	.89	-
Other application controls	Yes	Yes	-
Occupational fixed effects	Yes	Yes	Yes
N	5,636	5,662	3,536

Notes:

Table 3. Decomposition. Low and high skilled jobs.

	Ethnicity Experiment A		Gender Experiment B	
	Low skilled	High skilled	Low skilled	High skilled
A. Basic probit				
Arabic/Female name	-.098***	-.098***	.017	.037**
	[.010]	[.016]	[.012]	[.016]
B. Heteroskedastic probit				
Arabic/Female name (unbiased)	-.097***	-.100***	.021*	.039**
	[.011]	[.016]	[.012]	[.017]
Marginal effect of name through level	-.110***	-.064	-.025	.015
	[.040]	[.048]	[.035]	[.045]
Marginal effect of name through variance	.013	-.037	.046	.023
	[.034]	[.041]	[.035]	[.047]
Relative standard deviation of unobservables	1.05	.84	1.23	1.16
Wald test statistic, standard deviation == 1 (p-value)	0.71	0.32	0.23	0.64
Wald statistic, ratios of coefficients are equal (p-value)	0.77	0.93	0.34	0.19
Other application controls	Yes	Yes	Yes	Yes
Occupational fixed effects	Yes	Yes	Yes	Yes
N	3,533	2,103	3,549	2,113

Notes:

Table 4. Decomposition. Share of immigrants and females.

	Ethnicity Experiment A		Gender Experiment B	
	Share immigrants		Share females	
	above average	below average	above average	below average
A. Basic probit				
Arabic/Female name	-.079***	-.101***	.025**	.028**
	[.020]	[.010]	[.013]	[.014]
B. Heteroskedastic probit				
Arabic/Female name (unbiased)	-.084***	-.101***	.027**	.030**
	[.021]	[.010]	[.013]	[.014]
Marginal effect of name through level	-.065	-.098***	.008	-.113
	[.042]	[.042]	[.031]	[.212]
Marginal effect of name through variance	-.020	.004	.020	.149
	[.039]	[.036]	[.033]	[.159]
Relative standard deviation of unobservables	.92	.98	1.09	2.42
Wald test statistic, standard deviation == 1 (p-value)	0.60	0.92	0.57	0.53
Wald statistic, ratios of coefficients are equal (p-value)	0.95	0.72	0.72	0.66
Other application controls	Yes	Yes	Yes	Yes
Occupational fixed effects	Yes	Yes	Yes	Yes
N	1,043	4,593	2,963	2,699

Notes:

Table 5. Decomposition. Occupational demand.

	Experiment A		
	Occupational demand		
	Low	Medium	High
A. Basic probit			
Arabic name	-.058***	-.106***	-.106***
	[.014]	[.013]	[.017]
B. Heteroskedastic probit			
Arabic/Female name (unbiased)	-.064***	-.108***	-.104***
	[.015]	[.014]	[.017]
Marginal effect of name through level	.025	-.059	-.126***
	[.120]	[.107]	[.046]
Marginal effect of name through variance	-.089	-.050	.023
	[.134]	[.098]	[.035]
Relative standard deviation of unobservables	.67	.80	1.13
Wald test statistic, standard deviation == 1 (p-value)	.41	.56	.53
Wald statistic, ratios of coefficients are equal (p-value)	.97	.72	.99
Other application controls	Yes	Yes	Yes
Occupational fixed effects	Yes	Yes	Yes
N	1,329	2,509	1,798

Notes: Occupational demand is defined by the callback rate for female applicants in experiment B. The following jobs are defined as having a low demand: shop sales assistants, cleaners, and teachers (language)). The following jobs are defined as having a medium demand: mechanics, construction workers, accountants, high school teachers, business sales assistants, and teachers (science). The following jobs are defined as having a high demand: restaurant workers, motor vehicle drivers, computer professionals, and nurses. The three categories are based on whether the job specific callback rate for female applicants are <= 22 percentage points, > 22 percentage points but <= 32 percentage points, or > 32 percentage points.