Does the design of correspondence studies influence the measurement of discrimination?

Magnus Carlsson[1*], Luca Fumarco[1†] and Dan-Olof Rooth[1,2,3†]

* Correspondence:
Magnus.Carlsson@lnu.se
†Equal contributors
[1]Centre for Labour Market and
Discrimination Studies, Linnaeus
University, Kalmar, Växjö, Sweden
Full list of author information is
available at the end of the article

Abstract

Correspondence studies can identify the extent of discrimination in hiring as typically defined by the law, which includes discrimination against ethnic minorities and females. However, as Heckman and Siegelman (1993) show, if employers act upon a group difference in the variance of unobserved variables, this measure of discrimination may not be very informative. This issue has essentially been ignored in the empirical literature until the recent methodological development by Neumark (2012). We apply Neumark's method to a number of already published correspondence studies. We find the Heckman and Siegelman critique relevant for empirical work and give suggestions on how future correspondence studies may address this critique.

JEL classification: J71

Keywords: Correspondence studies; Discrimination

1. Introduction

Correspondence studies are an increasingly popular method for measuring discriminatory treatment against, e.g., ethnic minority and female workers in the labor market (see Riach and Rich, 2002, for a survey). In the standard correspondence study, matched pairs of qualitatively identical job applications are sent to employers that have advertised a job opening. The only difference between the fictitious applications is the name of the applicant, which signals ethnicity or gender. The degree of discrimination in hiring is quantified by calculating the difference in the callback rate (i.e., the fraction of invitations) to a job interview between the groups.

The advocates of correspondence studies argue that the method provides the most clear and convincing evidence of discrimination. Their main argument is that a carefully designed correspondence study can identify discriminatory treatment by employers since the signal of group belonging is randomized. This circumvents the problem with unobserved individual heterogeneity – a common problem in studies using administrative data.

The method's ability to identify discriminatory treatment by employers is certainly attractive, but it should be noted that correspondence studies cannot distinguish between preference-based (Becker, 1957) and statistical discrimination (Aigner and Cain, 1977; Arrow, 1973; Phelps, 1972). Somewhat simplified, preference-based discrimination is based on employer prejudice, while statistical discrimination arise when employers act

upon perceived group differences in the mean or variance of unobserved variables (i.e., variables not included in the job applications). However, the inability to separate between these two types of discrimination may not be a huge drawback unless the aim is solely to identify preference-based discrimination. In many countries, both preference-based and statistical discrimination against, e.g., ethnic minorities and women are illegal[1]. Hence, the level of discrimination being measured by the standard correspondence study is an unbiased measure of the degree of discrimination as defined by the law for these countries.

More problematic is that, despite being an unbiased measure of what the law defines as discrimination, which includes the case when employers act upon perceived group differences in the variance of unobserved variables, it may not be very informative. The problem is that when employers perceive a group difference in the variance of unobserved variables, the degree of discrimination in a correspondence study depends on the level at which the experimenter standardizes the observed characteristics in the job applications[2]. As a result, the standard correspondence study only tells the true level of discrimination against ethnic minority or female applicants who have similar qualifications as in the fictitious job applications. In the extreme case, a badly designed correspondence study may measure the degree of discrimination against a very atypical, or even non-existing, ethnic minority or female job applicant. In order to obtain an informative measure of discrimination in the market, the level of standardization must reflect the qualifications of a representative ethnic minority or female job applicant.

Although the idea of perceived group differences in the variance of unobservables has a long tradition in economics (e.g., Aigner and Cain, 1977), the issue has been essentially ignored in the empirical literature on correspondence studies until the appearance of the method proposed by Neumark (2012)[3]. In short, Neumark's method implies estimating the perceived relative variance in unobserved variables across groups, which then makes it possible to decompose the measured degree of discrimination into two parts. The first part captures discrimination in hiring due to employer preferences and/ or a perceived group difference in the mean of unobserved variables, while the second part captures discrimination in hiring due to a perceived group difference in the variance of unobserved variables. In our study, the second part is of main interest, since it reveals to what extent the result of a particular correspondence study is affected by its design, i.e., the level of standardization of the qualifications included in the job applications. Neumark applies his method to the data in the seminal correspondence study conducted by Bertrand and Mullainathan (2004) and finds indicative evidence of that the degree of discrimination depends on perceived group differences in the variance of unobserved variables. Baert et al. (2013) also applies Neumark's method, but to their own data, and find a similar result.

In the current study, we use Neumark's method to analyze to what extent a perceived group difference in the variance of unobserved variables is an issue in a number of already published correspondence studies. To this end, we use data from three experiments conducted in the Swedish labor market between 2005 and 2007. In two of the experiments, our results indicate that the degree of discrimination depends on perceived group differences in the variance of unobservables, while in one experiment there is no evidence of a dependency[4].

The next section explains the issue with perceived group differences in the variance of unobserved variables and the level of standardization (henceforth *the HS critique*, since it was first discussed in the seminal paper by Heckman and Siegelman, 1993). Section 3 explains Neumark's method in more detail, Section 4 describes the correspondence studies used to implement Neumark's method, Section 5 presents the main results, and Section 6 concludes.

2. The HS critique

This section aims at graphically explaining the intuition behind the HS critique. We first explain how an employer estimates the productivity of an applicant, and how an applicant's probability of being invited to a job interview is determined. Then we turn to the factors that determine the measured degree of discrimination in a correspondence study, where we focus on the level of standardization of the job applications. Much of the content of this section is inspired by Heckman and Siegelman (1993, henceforth HS)[5]. For readers that are interested in a more formal and detailed explanation we refer to HS's paper.

2.1 The productivity of a job applicant

To determine if a job applicant should be invited to a job interview, the employer estimates the productivity of the applicant. The productivity depends on deterministic observed variables X^{OBS} (variables that are included in the job application), unobserved random variables X^{UNOBS} (variables that are not included in the job application), and a discount factor γ that reflects employer preferences, which takes a negative value for applicants in the discriminated group and zero otherwise. Total productivity P for an applicant is then given by

$$P = \beta^{OBS} X^{OBS} + X^{UNOBS} + \gamma \tag{1}$$

where β^{OBS} is the return to observed characteristics and the return to unobserved characteristics has been normalized to one.

2.2 The probability of a job interview

Job applicants that have productivity above a certain threshold are invited to a job interview. Since the productivity of a particular job applicant is unknown, the employer estimates the likelihood that the productivity of a job applicant passes the threshold; the only random factor in the estimation is X^{UNOBS}. If X^{UNOBS} follows a normal distribution, total productivity P is also normally distributed[6]. The mean of P is $E[P] = \beta^{OBS} X^{OBS} + E[X^{UNOBS}] + \gamma$, which depends on the employer's perception about the mean of X^{UNOBS}, and the variance of P is determined by the employer's perception about the variance of X^{UNOBS}. Both the mean and variance of P may vary between groups. In Figure 1, the probability of being invited to a job interview is graphically illustrated. The shaded area is the probability of being invited to a job interview, which corresponds to the probability of passing the threshold c[7].

2.3 Discrimination

The measure of discrimination in a correspondence study reflects the situation where there are two groups of job applicants with identical observed characteristics X^{OBS}, but

$$E[P] = \beta^{OBS} X^{OBS} + E\left[X^{UNOBS}\right] + \gamma$$

Notes: The vertical axis measures density and the horizontal axis measures productivity. The shaded area is the probability of being invited to a job interview (i.e., the probability of passing the threshold c).

Figure 1 The probability of an invitation to a job interview.
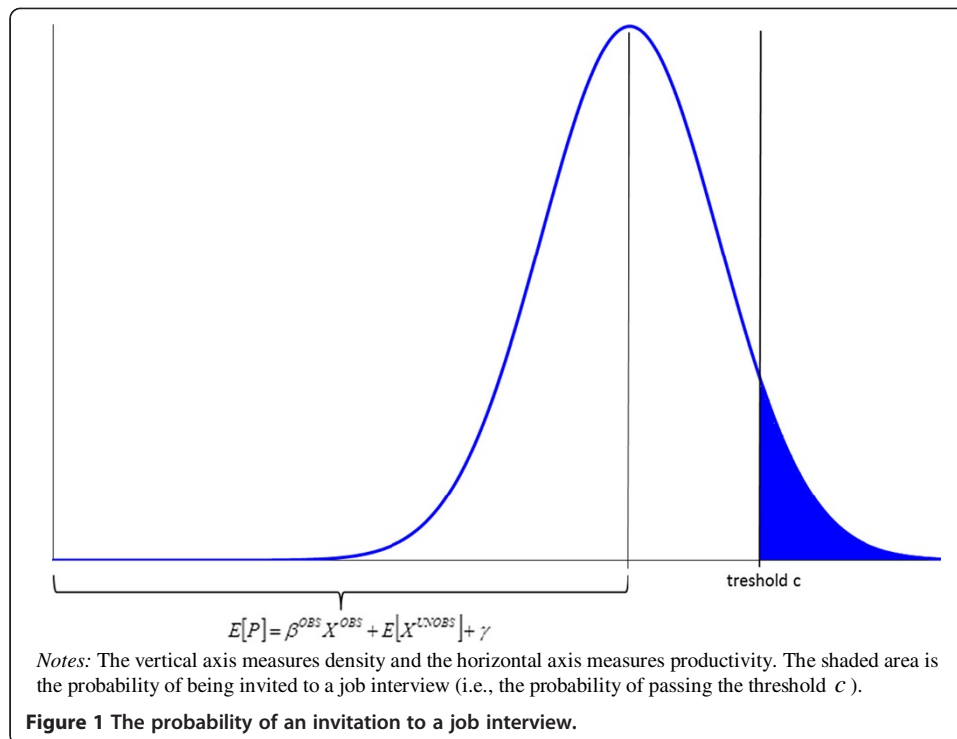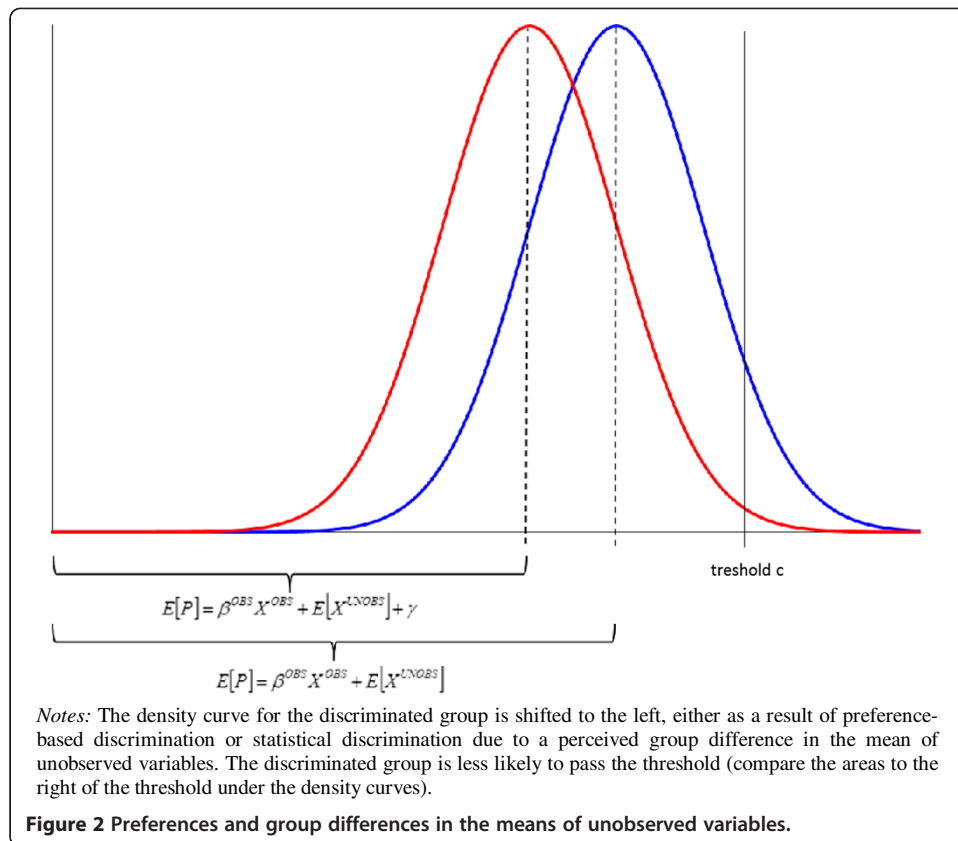
the likelihood of being invited to a job interview is higher for one group of applicants. This measure of discrimination can be decomposed into two parts. The first part, which we label *the effect through the level of discrimination*, reflects the combined effect on discrimination through preference-based discrimination and/or a perceived difference in the mean of unobserved variables. The second part, which we label *the effect through the variance of unobserved variables*, reflects the effect on discrimination through a perceived group difference in the variance of unobserved variables. The second part depends on the level of standardization, which we return to below.

The effect through the level of discrimination

Figure 2 illustrates the effect through the level of discrimination, i.e., a situation when there is no perceived group difference in the variance of unobserved variables. In this case, the expected productivity $E[P]$ is lower for applicants in the discriminated group for whom the density curve is shifted to the left. As a result, there is a lower probability of passing the threshold for these applicants. Note that $E[P]$ may be lower as a result of either preference-based discrimination ($\gamma < 0$) or statistical discrimination based on a perceived difference in the mean of unobserved variables ($\mathrm{E}[X^{UNOBS}]$). Hence, correspondence studies cannot distinguish between preference-based discrimination and statistical discrimination based on a perceived difference in the mean of unobserved variables.

The effect through the variance

A more problematic case, which is the focus of this paper, is statistical discrimination due to perceived group differences in the variance of unobserved variables. This type of discrimination is problematic, since its magnitude depends on the level of standardization

Notes: The density curve for the discriminated group is shifted to the left, either as a result of preference-based discrimination or statistical discrimination due to a perceived group difference in the mean of unobserved variables. The discriminated group is less likely to pass the threshold (compare the areas to the right of the threshold under the density curves).

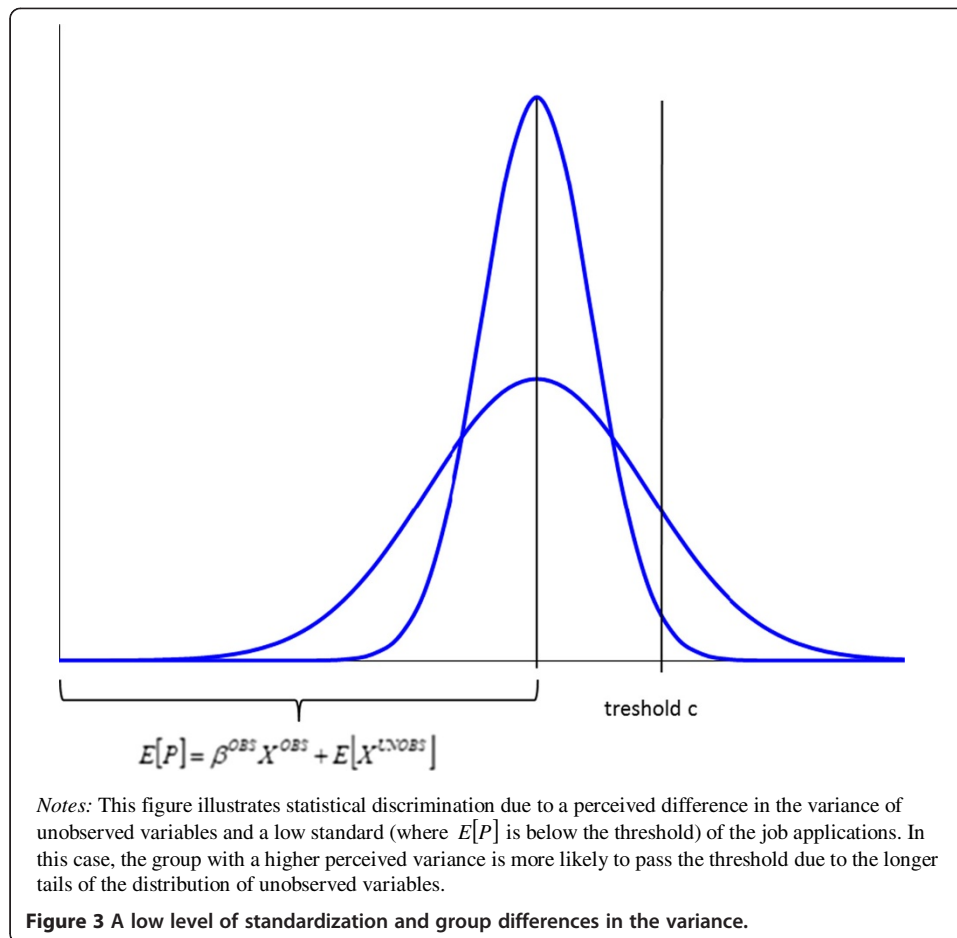**Figure 2 Preferences and group differences in the means of unobserved variables.**

of the job applications. To provide the intuition to this issue, let us focus on the simplest case where there are no perceived group differences in the mean of unobserved variables and no preference-based discrimination (i.e., $\gamma = 0$).

Figure 3 illustrates a situation where the experimenter has set a low standard of the job applications. In this case, the expected productivity $E[P]$ is below the threshold for both groups of applicants. However, if one group of applicants has a higher variance of unobserved variables, then job applicants from this group are more likely to pass the threshold due to the longer tails of the distribution of unobserved variables.

Figure 4 illustrates the opposite case where the experimenter has set a high standard of the job applications. In this case, the expected productivity $E[P]$ is above the threshold for both groups of applicants. However, in this scenario the group with the higher variance of unobserved variables will now be less likely to pass the threshold due to the longer tails of the distribution of unobserved variables.

The cases illustrated in Figures 3 and 4 give the theoretical argument to why the results from a standard correspondence study may not be very informative about the degree of discrimination in the market. The cases show that the measured degree of discrimination depends on the level of standardization of the job applications if there is a perceived group difference in the variance of unobserved variables. Hence, if the level of the qualifications is not set to mirror a representative ethnic minority or female job applicant, the measured degree of discrimination may say little about the average degree of discrimination in the market[8].
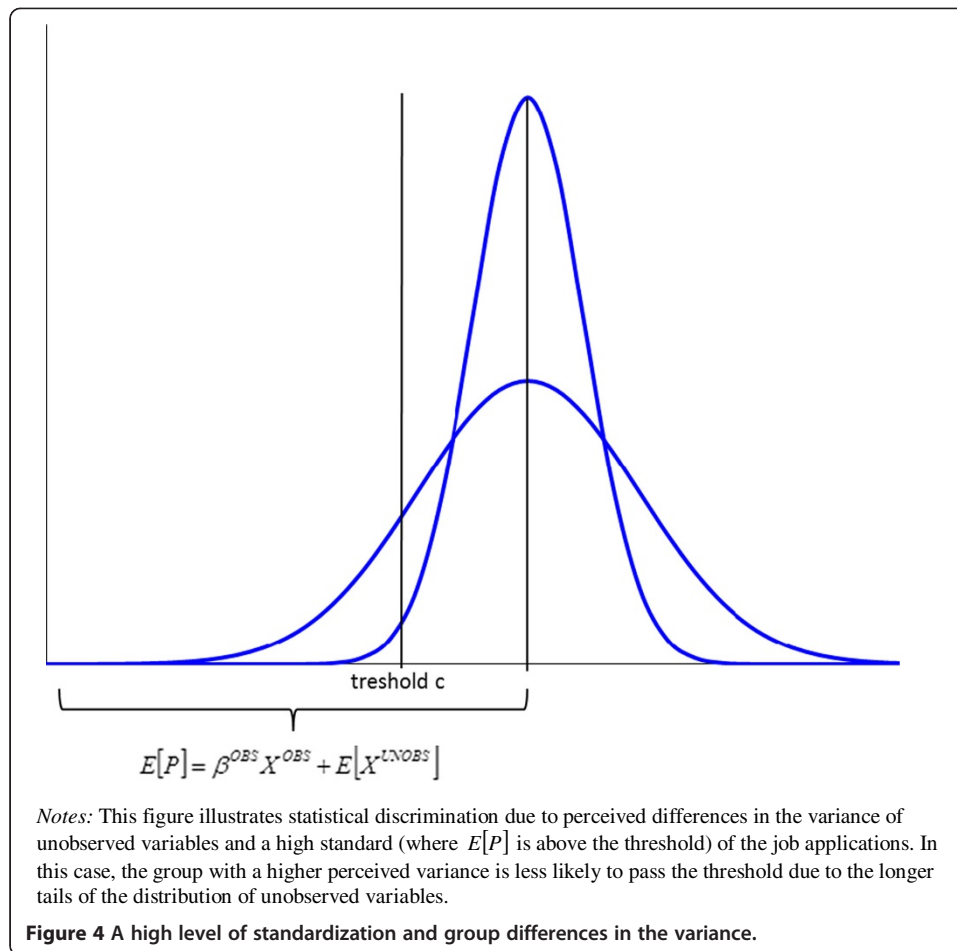
$$E[P] = \beta^{OBS} X^{OBS} + E\left[X^{LNOBS}\right]$$

Notes: This figure illustrates statistical discrimination due to a perceived difference in the variance of unobserved variables and a low standard (where $E[P]$ is below the threshold) of the job applications. In this case, the group with a higher perceived variance is more likely to pass the threshold due to the longer tails of the distribution of unobserved variables.

Figure 3 A low level of standardization and group differences in the variance.

3. Neumark's method[9]

Neumark's insight is that the HS critique can be addressed in a two step procedure. In the first step, the degree of discrimination is estimated together with the group specific variance of unobserved variables. In the second step, the estimated degree of discrimination is decomposed into two parts: the effect of group belonging *through the level of discrimination* (see Figure 2) and the effect of group belonging *through the variance of unobserved variables* (see Figures 3 and 4). In the analysis, we have followed Neumark's two step procedure, which we have implemented using Stata 12[10].

In the first step, Neumark uses the heteroskedastic probit model for estimation. Identification of the group specific variance in the heteroskedastic probit model requires data from a correspondence study that have random variation not only in the signal of group belonging, but also in some other observed productivity related variable(s) in the job applications. Importantly, there is an identifying assumption in the heteroskedastic probit model, which translates into an assumption of equal returns across groups to these additional productivity related variables. Below we return to whether this assumption is likely to hold.

In the second step, Neumark decomposes the marginal effect of group belonging in the heteroskedastic probit model into the two parts: the effect of group belonging *through the level of discrimination* and the effect of group belonging *through the*

$$E[P] = \beta^{OBS} X^{OBS} + E\left[X^{UNOBS}\right]$$

Notes: This figure illustrates statistical discrimination due to perceived differences in the variance of unobserved variables and a high standard (where $E[P]$ is above the threshold) of the job applications. In this case, the group with a higher perceived variance is less likely to pass the threshold due to the longer tails of the distribution of unobserved variables.

Figure 4 A high level of standardization and group differences in the variance.

variance of unobserved variables. The standard errors of the two parts are calculated using the delta method.

Returning to the identifying assumption of equal returns, this assumption is likely to hold for well designed correspondence studies, where there should be no group differences in the quality of the observed characteristics. E.g., in a written application the experimenter can easily choose not only the amount of schooling and work experience, but also similar schools and type of work experience so that the returns to those characteristics are the same across groups.

Moreover, the identifying assumption about equal coefficients can be tested. To implement the test, the first step is to estimate the probability of an invitation to a job interview separately for the two groups. In the second step, the residual standard deviations are normalized such that for one group the standard deviation is equal to unity while for the other group it is equal to the ratio of the group residual standard deviations. In the simplest case, with only one observed productivity related variable being varied in the job applications, a group difference in the coefficient of this variable can either arise because the identifying assumption does not hold or because the relative standard deviation is different from unity. However, with (at least) two observed productivity related variables that vary in the job applications, it becomes possible to test the null hypothesis of equal coefficients across groups of the observed applicant

characteristics. In the third step of the test, the ratios of the two coefficients are calculated separately for each group of applicants. It is the fact that the relative standard deviation cancels out for the second group, since this is a factor in both the dominator and numerator, which enables the test. In the final step, the null hypothesis of equal coefficients is tested by testing if the two ratios are equal across the groups[11]. We apply this test in our empirical analysis.

4. Data

To implement Neumark's method we use data from three different correspondence studies conducted in the Swedish labor market, which investigate both ethnic and gender discrimination, and have random variation in applicant characteristics. These three data sets are labeled Experiment A, B, and C. Recall that Neumark's method requires observed applicant characteristics that have a significant effect on the probability of an invitation and that the effect is the same across groups. Since the set of variables that fulfills this requirement vary across the experiments, we use a different set of observed characteristics for each experiment.

4.1 Experiment A

In Experiment A, focus is on ethnic discrimination against applicants with Middle Eastern sounding names and the data was collected in a field experiment conducted between March and November 2007[12]. 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. 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 include five variables that fulfill these requirements in the analysis of the variance of unobservables, while the other variables are excluded from the regressions.

The first two variables regard the personality of the candidate, basically following the Big Five taxonomy using the two of its five categories - agreeableness and extroversion (see Borghans et al. 2008). Being an *agreeable* person has both a moral and social dimension. An agreeable applicant states that it is important to care about others and likes to work in a group. In contrast, an applicant that is not agreeable does not emphasize these qualities[13]. Considering the category *extrovert*, it was decided in the design of the experiment to focus on the lower level category *competence*. A competent applicant states that he or she is a hardworking person that puts a lot of effort on the job. In contrast, an applicant that is not competent does not emphasize these qualities[14]. Both these variables are coded as dummy variables in the empirical analysis.

The third variable captures the type of neighborhood the applicant lives, with a dummy variable that indicates if the applicant lives in a high income area (i.e., mean income in the area is above the average). The fourth variable gives the applicant's previous work experience), 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 the applicant is engaged in sport activities or not. Sport activities could 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[15]. For these advertised jobs, 5,657 applications – 2,837 with a typical native Swedish sounding name and 2,820 with a typical Middle Eastern sounding name – were sent to 3,325 employers. Different job applications were used in each occupation in order to match the specific skills that are important in an occupation (this holds for Experiment B and C as well). 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.

4.2 Experiment B

Experiment B considers 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 in the same occupations. 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 construct 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 having good labor market characteristics if he or she has at least one 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. All other explanatory variables are excluded from the regressions.

4.3 Experiment C

Also Experiment C considers ethnic discrimination against applicants with Middle Eastern sounding names. Actually, Experiment C consists of observations from two different correspondence studies[16]. What justifies viewing them as a single experiment 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 are born in Sweden, have either a typical native Swedish or Middle Eastern sounding name, are on average 25–30 years old, have two to four years of work experience in the relevant occupation, and have obtained their education in the same type of school. Also, in both studies, the applications consist of a quite general biography on the first page and a detailed CV of education and work experience on the second page. Finally, in both studies 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.

Despite the similarities between these two studies there is one important distinguishing factor. 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; the quality of the applications in terms of labor market experience and skills were raised in three occupations and lowered in the other three occupations[17]. These six occupations contain 3,536 observations. This calibration generates the variation in the standards of the applications that we utilize in the current paper. However, since only one variable varies in Experiment C, we are not able to test the identifying assumption of equal returns to characteristics in this case.

5. Empirical analysis

In this section, we first use the standard probit model to provide a set of basic results for Experiment A-C. Then we turn to the main analysis, where we apply Neumark's method.

5.1 Basic results

Table 1 presents basic results for Experiment A-C. The purpose of this table is to report the estimate of the degree of discrimination that comes out of a standard correspondence study and to show that the observed variables in the job applications that we use to implement Neumark's method have significant effects – with expected signs – on the probability of receiving an invitation to a job interview[18].

Experiment A

The first two columns of Table 1 report the basic results for Experiment A. In the top row of the first column, we find that the ethnic difference in the probability of a job interview is 9.4 percentage points. From the following seven rows of this column it is evident that applicants that are extrovert, agreeable, live in a high income area, 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 sign – 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 all application attributes and occupational fixed effects. These additional controls include dummy indicators for whether the vacancy 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.

Experiment B

The basic results for Experiment B are found in the third and fourth columns of Table 1. Again, in the top row we find the group difference in the probability of a job interview, 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.

Table 1 Basic probit results

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

Notes: The dependent variable is an indicator variable that takes the value one if the applicant was invited to a job interview and zero otherwise. The reported standard errors (in brackets) are clustered at the job advertisement level. ***, **, and * denote the 1, 5 and 10 percent significance levels, respectively.

Experiment C

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 sounding names. The row at the bottom of the table reveals that improved quality applications have a significantly higher probability of a job interview. Note that there is only one column of 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 the high quality variable. 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 are manipulated at the occupational level, which means that without occupational fixed effects the estimate of improved quality may also reflect occupation specific demand.

5.2 Main results

In this section, we use Neumark's method where we first estimate a heteroskedastic probit model, and then decompose the estimated degree of discrimination into two parts: the effect of group belonging through *the level of discrimination* and through *the variance of unobserved variables*. An issue is that Neumark's method often increases the standard errors of the decomposed marginal effects by a factor of 2.5, or more, compared to the standard error of the undecomposed marginal effect, which renders statistically insignificant estimates. Since we find the magnitude of the estimates to be

economically important we choose to view the decomposed marginal effects as still providing evidence.

To facilitate the interpretation of the results of the decomposition, we discuss the results for the first experiment (Experiment A) in detail, while the results of the remaining experiments are discussed more briefly.

Experiment A

The results of the decomposition for Experiment A (ethnic discrimination) are presented in the first column of Table 2. The estimate in panel A is for comparison and repeats the standard probit estimates of having a typical Middle Eastern sounding name in Table 1, while the estimate in panel B presents the estimated marginal effect obtained from the heteroskedastic probit model. If there is a group difference in the variance of unobservables, this would violate the standard probit model assumption of equal variances. Therefore, a first indication of a difference in the variance of unobservables would be if the estimate of the degree of discrimination differs between the standard probit and the heteroskedastic probit. However, in the case of Experiment A, the two estimates are very similar, which indicates that there is no perceived group difference in the variance of unobservables.

Next, the first two rows in panel C give the marginal effects of group belonging decomposed into *the effect through the level of discrimination* and *the effect through the variance of unobserved variables*, respectively. The key estimate of interest is *the effect through the variance*, since this estimate tells if the degree of discrimination depends on perceived group differences in the variance of unobserved variables. In Experiment A, the point estimate of *the effect through the variance* is small and insignificant. This implies that the estimate of discrimination in this experiment does not depend on the design of the experiment, i.e., the level of standardization being set by the experimenter.

Table 2 Decomposition

	Ethnicity Experiment A	Gender Experiment B	Ethnicity Experiment C
	(1)	(2)	(3)
A. Basic probit			
Middle Eastern sounding/female name	-.096*** [.009]	.026** [.009]	-.128*** [.010]
B. Heteroskedastic probit			
Middle Eastern sounding/female name	-.098*** [.009]	.029*** [.010]	-.135*** [.011]
C. Decomposition			
Marginal effect of name through level	-.088*** [.023]	.001 [.024]	-.090** [.029]
Marginal effect of name through variance	-.010 [.025]	.028 [.025]	-.044 [.033]
Relative standard deviation of unobserved variables	.96	1.13	.83
Wald test statistic, standard deviation == 1 (p-value)	.68	.29	.14
Wald statistic, ratios of coefficients are equal (p-value)	.67	.89	-
Other application controls	Yes	Yes	-
Occupational fixed effects	Yes	Yes	Yes
Number of observations	5,636	5,662	3,536

Notes: The dependent variable is an indicator variable that takes the value one if the applicant was invited to a job interview and zero otherwise. The reported standard errors (in brackets) are clustered at the job advertisement level. ***, **, and * denote the 1, 5 and 10 percent significance levels, respectively.

The next two rows (in panel C), respectively, present the point estimate of the relative standard deviation of the unobservables for applicants with typical native Swedish and Middle Eastern sounding names and the resulting p-value from testing the hypothesis that the relative standard deviation equals one. There is no evidence of a perceived difference in the variance of unobserved variables, since the relative standard deviation is close to one (.96) and the p-value is large (.68).

The last row in panel C contains the result from testing the identifying assumption of equal coefficients across groups for the observed applicant characteristics. The high p-value for the Wald statistic suggests that the data is consistent with the identifying assumption of equal coefficients.

Experiment B

The results of the decomposition for Experiment B (gender discrimination) are presented in the second column of Table 2. Here, the difference between the two estimates from the standard probit and the heteroskedastic probit is larger – at least in relative terms (compare the estimates in panel A and B). This indicates the existence of a perceived difference in the variance of unobserved variables across groups. Interestingly, when the estimated degree of discrimination is decomposed, we find that *the effect through the variance* is of the same magnitude, although insignificant, as the overall estimate of discrimination (second estimate in panel C), while *the effect through the level* is zero (see first estimate in panel C). This suggests that the measured degree of discrimination in this experiment depends on the level of the standardization of the job applications being set by the experimenter.

Although statistically insignificant, if we take the estimate of the relative standard deviation of unobservables at face value, the interpretation 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 (where $E[P]$ is below the threshold, see Figure 3) of the applications being set in the experiment and where the higher variance of unobserved characteristics benefits females.

The high p-value for the Wald statistic in the last row in panel C suggests that the data is consistent with the identifying assumption of equal coefficients for the observed applicant characteristics.

Experiment C

The results of the decomposition for Experiment C (ethnic discrimination) are presented in the third column of Table 2. Here, the difference between the two estimates from the standard probit and the heteroskedastic probit is larger than in both Experiment A and B, which suggests that employers have acted on perceived group differences in the variance of unobservables when hiring. As expected, when the estimated degree of discrimination is decomposed, *the effect through the variance* is quite large, but statistically insignificant. Taking the point estimate at face value suggests that the level of standardization plays a role for the estimated degree of discrimination in this experiment.

Although the relative standard deviation of unobservables 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 Middle Eastern sounding names is only .83 of the standard deviation for applicants with native Swedish sounding names. Similarly as in Experiment B, this is consistent with setting a low standard (where $E[P]$ is below the threshold, see Figure 3) of the applications in the experiment where applicants with typical Middle Eastern sounding names are suffering from their lower variance of the unobserved variables.

In this experiment, it is not possible to test the identifying assumption of equal returns to observed applicant characteristics, since, in addition to the Middle Eastern sounding name dummy, there is only variation in one explanatory variable.

6. Concluding remark

It can be argued that correspondence studies provide the most clear and convincing evidence of discrimination since the signal of group belonging in these studies is randomized, which circumvents the problem with unobserved individual heterogeneity. However, the results in HS show that the measured degree of discrimination in a correspondence study may still not be very informative if the level of qualifications of the fictitious job applications do not match up with the representative job applicant in the discriminated group. The reason is that when employers act upon perceived group differences in the variance of unobserved variables the degree of discrimination depends on how the experimenter sets the level of qualifications in the job applications. This so called HS critique has essentially been ignored in the empirical literature on correspondence studies until the appearance of the methodology proposed by Neumark (2012).

We use Neumark's method to reexamine a number of already published standard correspondence studies, which do not take into account the level of standardization. We find suggestive evidence that the results of discrimination depend on the level of standardization and hence, perceived group differences in the variance of unobserved variables may be important, not just as a theoretical argument, but also for the empirical design when conducting correspondence studies.

What are the implications of our findings? We believe our results are sufficiently strong to suggest that correspondence studies cannot continue to ignore the issue raised by HS. In our opinion, future correspondence studies should try to invest more effort in the design of the job applications, aiming for a level of standardization that reflects the representative ethnic minority or female job applicant in the population (and ideally for each type of job). This requires information on what qualifications real job applicants have and a challenge may be that, at least historically, such information has been difficult to obtain. However, today there exist large databases with such information as a result of online job search: Job applicants put their CVs online to make them available for employers searching for workers. We believe that a natural way forward would be to use information from such databases when designing the job applications to obtain a truly informative measure of the degree of discrimination in the labor market. Also, we believe that future correspondence studies should be designed to satisfy the requirements necessary to implement Neumark's method, which makes it possible to, at least in retrospect, analyze to what extent the measured degree of discrimination depends on the level of standardization.

Endnotes

[1]Under European law, which applies to the member countries of the European Union, of which Sweden is a member, discrimination in employment situations based on, e.g., nationality, race, ethnic origin, and gender is considered a crime. Discrimination under European law includes both preference-based and statistical discrimination in employment situations by covering general situations where "one person is treated less favorably in a comparable situation" (European Union Agency for Fundamental Rights, 2011). Similar legislation is found in many other countries, including the U.S. (Riach and Rich, 2002).

[2]This idea was originally formulated by Heckman and Siegelman (1993) stating that, if perceived group differences in unobserved variables exists, *preference based* discrimination is unidentified in a correspondence study. Heckman (1998) also discusses this issue.

[3]This issue is also discussed in Neumark (2013), but his method is applied in Neumark (2012).

[4]These experiments are found in Carlsson and Rooth (2007), Carlsson (2010), Rooth (2010), and Eriksson and Rooth (2014).

[5]See also Heckman (1998).

[6]As HS argue, the results shown here hold for all distributions in the family of bell shaped distributions.

[7]A subtle issue, which Neumark (2012) points out, is that it should actually be deterministic who is invited to a job interview and who is not, if all employers make the same probability calculation and invite applicants based on the probability of passing the threshold. Obviously, this is not the pattern we see in reality. However, it is straightforward to incorporate a random component into the framework that describes the employers' decision making. One way is to assume firm specific thresholds that are, e.g., normally distributed.

[8]Note that nothing essential changes in our conclusions if we also allow for preference-based discrimination and/or perceived differences in the mean of unobserved variables. This would only affect the probability to be hired for the discriminated group, or both groups, either counteracting or reinforcing the effect from the level of standardization through perceived differences in the variance.

[9]Much of the content of this section is taken from Neumark (2012). For a more detailed explanation of the issues involved in this section the reader should turn to Neumark (2012).

[10]The Stata code is available upon request.

[11]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.

[12]Details of this experiment are found in Eriksson and Rooth (2014), Carlsson and Rooth (2012), and Rooth (2011).

[13]The text (translated from Swedish) for *agreeable* is "My friends and former colleagues would probably state that I am a warm and social person who gets along great with others. Also, I think it is important to ensure people's needs, and not just focusing on the economic side. I have a strong empathy with people who are less fortunate than

myself and I am active in the Red Cross relief work", while the text for the opposite is "I usually do not sit and keep my opinions to myself but rather instead say what I think. Some of my former colleagues would probably call me a bit stubborn, but I believe it is important to be correctly understood and to get the job done. I do not mind working alone, since it is then sometimes easier to concentrate on the job task".

[14]The text for *competence* is "I am used to put great effort into work and I always try to do my best. I strive to be as precise as possible so the work tasks need not to be repeated. My old work mates would probably say that I am a person who always manage to get the job done. In addition, I would describe myself as a hardworking and tenacious (sw: uthållig) person who withstand stress", while for the opposite it is "I really like to work but at the same time I think it is important to keep a balance between work and leisure. The best days are the ones when I feel I have done my job and yet have energy to be active in my spare time. It is not important for me to be the best at work and my colleagues would probably describe me as a pretty relaxed".

[15]The included occupations were accountants, business sales assistants, cleaners, computer professionals, construction workers, language teachers in upper compulsory school, math/science teachers in upper compulsory school, mechanics, motor-vehicle drivers, nurses, restaurant workers, shop sales assistants, and teachers in secondary school.

[16]The first one is Carlsson and Rooth (2007) and the second one is Carlsson (2010).

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

[18]The estimates in Table 1 are obtained using the dprobit command in Stata 12.

Author details

[1]Centre for Labour Market and Discrimination Studies, Linnaeus University, Kalmar, Växjö, Sweden. [2]IZA, Bonn, Germany. [3]CReAM, London, UK.

References

Aigner DJ, Cain GG (1977) Statistical theories of discrimination in labor markets. Ind Labor Relat Rev 30(2):175–187
Arrow K (1973) The theory of discrimination. In: Ashenfelter O, Rees A (eds) Discrimination in labor markets. Princeton University Press, Princeton
Baert S, Cockx B, Gheyle N, Vandamme C, Omey E (2013) Do Employers Discriminate Less if Vacancies are Difficult to Fill? Evidence from a Field Experiment. Discussion Paper Series #7145, IZA, Bonn
Becker GS (1957) The economics of discrimination. University of Chicago press, Chicago
Bertrand M, Mullainathan S (2004) Are Emily and Greg More Employable than Lakisha and Jamal? A field experiment on labor market discrimination. Am Econ Rev 94(4):991–1013
Borghans L, Duckworth AL, Heckman JJ, Ter Weel B (2008) The economics and psychology of personality traits. J Hum Resour 43(4):972–1059
Carlsson M (2010) Experimental evidence of discrimination in the hiring of first-and second-generation immigrants. Labour 24(3):263–278
Carlsson M, Rooth DO (2007) Evidence of ethnic discrimination in the Swedish labor market using experimental data. Labour Econ 14(4):716–729

Carlsson M, Rooth DO (2012) Revealing taste-based discrimination in hiring: a correspondence testing experiment with geographic variation. Appl Econ Lett 19(18):1861–1864

Eriksson S, Rooth D-O (2014) Do employers use unemployment as a sorting criterion when hiring? Evidence from a field experiment. Am Econ Rev 104:1014–1039

European Union Agency for Fundamental Rights (2011) Council of Europe, Handbook on European non-discrimination law., http://www.echr.coe.int/Documents/Handbook_non_discri_law_ENG_01.pdf. (Accessed 24 February 2014)

Heckman JJ (1998) Detecting discrimination. J Econ Perspect 12(2):101–116

Heckman JJ, Siegelman P (1993) The Urban Institute audit studies: Their methods and findings. In: Fix M, Struyk R (eds) Clear and convincing evidence: Measurement of discrimination in America. Urban Institute Press, Washington DC, pp. 187-258

Neumark D (2012) Detecting discrimination with audit and correspondence studies. J Hum Resour 47(4):1128–1157

Neumark D (2013) Ethnic Hiring. In: Constant AF, Zimmermann KF (eds) International Handbook on the Economics of Migration. Edward Elgar, Cheltenham, UK, and Northampton, USA, pp 193–213

Phelps ES (1972) The statistical theory of racism and sexism. Am Econ Rev 62(4):659–661

Riach PA, Rich J (2002) Field experiments of discrimination in the market place. Econ J 112(482):F480–F518

Rooth DO (2010) Automatic associations and discrimination in hiring: real world evidence. Labour Econ 17(3):523–534

Rooth DO (2011) Work out or out of work — the labor market return to physical fitness and leisure sports activities. Labour Econ 18(3):399–409