What can we learn from correspondence testing studies?
What can we learn from correspondence testing studies?

Magnus Carlsson

Dan-Olof Rooth†

Abstract: Antidiscrimination policies play an important role in public discussions. However, identifying discriminatory practices in the labor market is not an easy task. Correspondence testing provides a credible way to reveal discrimination in hiring and provide hard facts for policies. What is this instrument? What does it show and how reliable is it? Should it be widely used for policymaking? Answers to these questions are provided.
“Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.”

Charter of Fundamental Rights of the European Union, Article 21 on Non-discrimination, Paragraph 1

1. Introduction

The EU legislation prohibits discrimination in employment based on race or ethnicity, gender, age, disability and religion or beliefs. However, there are reasons to expect that discriminatory behavior still exists among employers. Attitude surveys of the general public show evidence of negative attitudes toward minority groups and surveys among potentially discriminated groups also point in this direction for a number of EU countries. For example, for Sweden attitude surveys among the general public and surveys of the minority group indicate that ethnic discrimination is worst against individuals with a Middle Eastern background. Also, unemployment rates for immigrants born in the Middle East are found to be several times higher compared to native ones, indicating that ethnic discrimination exists in the recruitment process.

However, discrimination is just one possible explanation for group differences in finding jobs. Another is (to the researcher) unobserved productivity characteristics, such as language skills in the case of identifying ethnic discrimination. Since researchers seldom can account for all differences in productivity characteristics between groups it is a difficult task to empirically identify the extent of discrimination in employment using a standard regression approach (see Altonji and Blank, 1999).

To circumvent this difficulty researchers have relied on using field experiments specifically designed to test for discrimination in recruitment. The standard correspondence study sends matched pairs of qualitatively identical applications to employers that have advertised a job opening, the only difference being the signal of group belonging, like the name in the study of ethnic discrimination. 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 the correspondence testing method provide the most clear and convincing evidence of discrimination. Is this true? In the remainder of this article we will give a background of the correspondence testing method, discussing how to interpret the raw data and its implications, and finally also discuss its use in policymaking combating discrimination in employment. Although the focus in this article will
solely be on measuring discrimination on hiring using the correspondence testing experiment, the discussion could easily be extended to measuring discrimination in the housing market.

2. Introduction

Some forty years ago Jowell and Prescott-Clarke (1970) were the first to use the correspondence testing method to detect discriminatory treatment in hiring. They measured if employers in Birmingham more often gave an invitation to an interview to majority (White) than a minority (Asian/West Indian) job applicant in white collar jobs. Applying to thirty-two job vacancies per minority group they found a difference in the callback rate for a job interview of forty and nine percentage points for Asian and West Indian applicants, respectively, compared to majority applicants. Today, with the internet revolution and much recruiting occurring by email, the correspondence testing experimental method has been facilitated to amount to sending thousands of job applications to advertised job openings. For instance, Oreopoulos (2011) sent over thirteen thousand resumes in his study of ethnic discrimination in Canada. This development has made the correspondence studies an increasingly popular method for measuring discrimination in the labor market.

Although most correspondence testing studies belong to the category where the name is used as a signal of group belonging, hence, studying gender and ethnic discrimination, recent advances also study age, disability, sexual orientation and appearance/looks. Compared to using names, the empirical design to signal group belonging becomes a lot more challenging in these cases. The problem arises since, in the case of age, altering age one usually also alters experience. For the other three cases, the employers might view the applications as odd if the applicant signals he/she is gay/lesbian, disabled or presents a picture of him/herself in a job application (signaling a particular appearance). Hence, it might in this case be problematic to view any difference in callbacks across groups as definitely arising out of discriminatory acts.

Correspondence testing studies have been implemented across a wide array of countries and demographic groups and they find evidence of discrimination in hiring against ethnic groups, and women in both the US, Canada and various EU countries (Neumark, 1996; Riach and Rich, 2002; Bertrand and Mullainathan, 2004; Pager 2007; Oreopoulos, 2011), by age and disability (Riach and Rich, 2002; Lahey, 2008; Ahmed, Andersson and Hammarstedt, 2012), by sexual orientation (Ahmed, Andersson and Hammarstedt, 2013) and appearance (Rooth, 2009). In addition, both UK and U.S. courts allow organizations that conduct correspondence studies to file claims of discrimination based on this evidence (Neumark, 2012). There are
also a number of studies that use correspondence testing to measure discrimination in the housing market (Ahmed and Hammarstedt, 2008).

3. Interpretation of results from a CT experiment

The interpretation issues involved when reading the results from a correspondence testing experiment are illustrated with the results from Carlsson and Rooth (2007). In preparing the job applications an important part is the choice of which observable productivity-related characteristics to standardize the applications on. As will be returned to below, depending on these choices the design will potentially impact on the interpretation of the results. The aim for any correspondence testing study must be to include the job-specific productivity characteristics being most important for hiring.

The last row of Table 1 gives the aggregated results of their experiment. From the first column it is evident that the two applications were sent to 1,552 different job openings. In 1,030 cases neither application was invited for interview and in the remaining 522 cases at least one of the two applications was invited for interview. Both applications were invited in 239 cases, while only the majority applicant was invited in 217 cases and only the minority applicant in 66 cases. From this information we can then calculate the probability to receive a callback for a job interview for majority and minority applicants, which is 29 and 20 percent, respectively. These numbers are arrived at by adding the number of times both groups received a callback for interview (Equal treatment) and the number of times only either group received a callback (Only majority/minority invited), divided by the number of jobs applied to.

The design of the experiment insures that these differences in callback rates between majority and minority applicants is due to the fact that firms/recruiters use group belonging as a decision variable in the selection process on who to call for interview. One can report this difference in relative or absolute terms. In relative terms we find that the probability of receiving a callback for a job interview is fifty percent lower for minority applicants compared to majority applicants. In absolute terms the differences in callback rates across groups is nine percentage points.

We have now arrived at an average measure of discriminatory treatment by the employers being part of the experiment, but how can we interpret this estimate and how can we use this result outside of the experiment, for instance, to make suggestions for policy? Before we turn to these issues it should be stated that correspondence testing only measures discrimination at the first stage in the hiring process. It is not aimed at capturing unequal
treatment in who gets the job, in promotions or in wage growth. For studying those dimensions other methods have to be applied (see for example Altonji and Pierret, 1997).

Employer preferences or statistical discrimination?
For policy purposes one would like to be able to identify whether the difference in callbacks across groups in a correspondence testing study arises from preference/taste based discrimination or from statistical discrimination. Correspondence testing studies ideally attempts to measure employer preferences/tastes for hiring majority over minority job applicants by controlling for the most important productivity related characteristics (see the appendix for a template). However, Heckman (1998) shows that unless the correspondence testing study includes all characteristics being important for hiring, and which differs on average across groups, the correspondence testing method cannot separately identify the mechanisms that drive discriminatory treatment. Hence, although that a carefully designed correspondence study potentially should allow the researcher to circumvent the problem with unobserved individual heterogeneity there are uncertainties regarding interpreting the group difference in callback rates for job interviews solely as arising from taste based discrimination.

However, this criticism should not be exaggerated since it only impacts on the interpretation of the discrimination. It might be argued 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. Since recruiters are not allowed (by discrimination legislation) to use or to make assumptions about group differences in characteristics not included in the job applications, any role that they have in the hiring process for the correspondence testing experiment could be interpreted as statistical discrimination. Hence, the group difference in callback rates in a correspondence testing study can be interpreted as capturing the combined effects of taste based discrimination and statistical discrimination.

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. The design of the correspondence testing experiment is in this respect important since the richer the set of applicant characteristics the less likely it becomes that statistical discrimination plays much of a role for group differences in hiring.

Is this the discrimination being observed in the market?
Most correspondence testing experiments respond to job adds being posted by firms in newspapers or on internet sites. Unfortunately those firms are most likely not a random sample of all firms in the market and the credibility of the experiment relies on to what extent the authors are able to provide facts about the firms and the channels they use to search for workers. For instance, it could be the case that only firms that are less discriminatory use channels which everyone could use, like newspaper adds, and hence, discrimination of an average firm employing labor is underestimated. Although no correspondence testing experiment has been able to arrive at a random sample of employers, some studies have collected basically all jobs being posted within a year in certain occupations. Hence, it is probably fair to say that if a minority worker would use these channels for their job search this is the level of discrimination they would encounter. Further, additional evidence on which channels the minority group under study uses when searching for jobs helps in evaluating the importance of the studies finding.

Hence, the firms being part of the correspondence testing experiment might not be representative of the firms that minority workers search for jobs at. It could actually be the case that many employers have a preference against hiring minority workers but since these are never approached by minority workers discrimination does not have an effect on the probability of finding jobs. Heckman (1998, pp. 102-103) puts it in this way: “The impact of market discrimination is not determined by the most discriminatory practices in the market, or even by the average level of discrimination among firms, but rather by the level of discrimination at the firms where ethnic minorities or women actually end up buying, working and borrowing. It is at the margin that economic values are set……Purposive sorting within markets eliminates the worst forms of discrimination”.

However, politicians might still be interested in if discrimination exists towards a certain minority group in particular occupations before it proposes a policy to change group ratios in employment. For instance, many countries advocate more balanced gender ratios in the labor market and it might interesting to know whether barriers to such a policy exists and in what occupations before balancing policies are implemented.

Choice of occupations and external validity
The relevance of the correspondence testing experiment in terms of its external validity increases if more occupations important for the labor market are included. The objective when choosing which occupations and regions to include in an experiment is to get a representative picture of the labor market, while at the same time designing a study that would be feasible to
implement in practice. To get a representative picture of the labor market, one would like an experiment to include a variation in the skill level of the job since there could be important differences in discrimination depending on the skill level of the job. Ideally, the experimenter should report the shares of total employment or total vacancies made up by the occupations included in the experiment.

However, it is also important to include many occupations to get a picture of how discrimination varies by occupation. Returning to Table 1 it gives the same type of data description as above, but now for high and medium/low skill occupations separately. If this study only had included the medium/low skilled occupations the conclusion would have been that discrimination is much more severe than if only the high skilled occupations were included.

A recent advance in the empirical design of the correspondence testing experiment makes it possible to benchmark the level of discrimination being found to, for instance, the return to job experience being estimated in the same experiment. This methodology implies that the researcher does not only randomly vary group belonging but also job experience, see Bertrand and Mullainathan (2004) and Rooth (2011). This benchmarking makes it possible to ask questions like “How does gender discrimination in hiring relate to the return to one extra year of experience?”.

An additional identifying issue
Above it was discussed to what extent a CT experiment is able to separate taste based discrimination from statistical discrimination when not all characteristics important for the hiring decision, and for which the groups being studied differ on average, are being included in the CT experimental design. In this section we discuss yet another type of identification problem being related to group differences in the variance of unobserved – or left out - characteristics, potentially having the consequence of not even being able to identify the level of discrimination. Heckman (1998) convincingly show that correspondence studies can obtain biased estimates of discrimination – in any direction – if employers evaluate applications according to some threshold level of productivity. The source of this bias in 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. This issue has essentially been ignored in the empirical literature on correspondence testing experiments until the appearance of the methodology proposed by Neumark (2012). Neumark applies his method on the correspondence testing data used in Bertrand & Mullainathan (2004) and finds suggestive evidence that the estimated degree of discrimination in the original paper is slightly underestimated. Future studies using the Neumark method will reveal to what extent this criticism of the correspondence testing method is empirically justified.

*Should correspondence testing be used to identify discriminating employers?*

An implication from the design of sending two otherwise identical applications with only group belonging being the difference across them to one single firm is that the outcome could potentially be used to detect discriminating employers and/or provide evidence whether a specific employer discriminates when hiring. As previously mentioned, the correspondence testing method is allowed as evidence in court in both the UK and the US. However, the results from the experiment in Carlsson and Rooth indicate that this should not be the sole evidence. The question to be asked is if the correspondence testing method can be used to state if a particular firm truly discriminated? In other words, is the correspondence testing experimental procedure useful for identifying a single firm consciously choosing one applicant over the other. In the Carlsson and Rooth study, only the minority applicant receives a callback sixty-six times (4 percent). Why that happens could have several explanations. First, there could simply exist employers having a preference for hiring the minority applicant. However, a look at the data show that all sixty-six of these recruiters have a majority background so even if this could be true, alternative explanations exists. The applications are sent in random order and it could be the case that only the application which arrives first receives a callback. However, in some of these cases the minority application was sent last and still received a callback. Hence, it could also be the case that the employer/recruiter had a miserable day the day the majority application arrived and perhaps missed out on it completely, or evaluated everything that day to the worst. What this shows is that the results using this experimental design probably also incorporate some randomness at the firm level, which cast doubts about the result from a correspondence testing being used as single evidence in a court case. However, it should also be said that this randomness plays a minor role for interpreting the average level of discrimination in the correspondence testing experiment since this randomness is affecting both groups equally.
Can we learn something about those who discriminate?

As is already evident firms are not randomly selected and hence it becomes problematic to state the effects of certain recruiter and/or firm attributes on discriminatory practices when hiring using the correspondence testing experimental design. For instance, in Carlsson and Rooth (2007) they add information on gender for the person being responsible for the hiring and the size of the firms and find that discriminatory practice is mainly a male phenomena which occurs in small firms. However, this result might have little to do with these attributes and hence, discrimination will probably not disappear if all male recruiters where to be replaced by females. It could be the case that the females in this sample of firms happen to be less discriminatory, perhaps from being more experienced with recruiting. There are also studies which attempt to measure the attitudes of the recruiters being part of the correspondence testing experiment. Rooth (2011) finds that the probability to invite Arab-Muslim job applicants decreases significantly when the recruiter responsible for the hiring has stronger negative implicit associations toward Arab-Muslim men. This suggests that automatic processes may exert a significant impact on employers’ hiring decisions. However, his result is subject to the same skepticism as above since the recruiters could be selective on other dimensions being responsible for the result.

Although that these additions to the experiment do not produce causal effects the results could still be useful input when designing more controlled laboratory experiment to learn why employers discriminate when hiring.

Ethical concerns

In a correspondence testing experiment employers are approached by fictitious job applicants who do not want employment. Obviously these employers are deceived and not asked about their consent to participate in the experiment. The discussions on ethics for correspondence testing studies therefore evolve around the issue of deception and the absence of informed consent. Rich and Riach (2004, p.459) argue that “no harm results from labour market field experiments, because individuals are not identified on publication, and inconvenience to employers and genuine applicants is minimised by offers of interview or employment being promptly declined.”. In addition, it is argued “that there can be no legitimate expectation of privacy in the act of hiring labour, as national governments and international bodies have accepted the onus of ensuring equality of opportunity for all citizens by declaring discrimination in employment unlawful.” (see also Pagan, 2007). There is also an opposing
view that non-deceptive practices is a public good and if researchers would extensively use deception this might change subject behavior and make experiments harder to interpret. For instance, employers might refrain from posting their job adds in newspapers and instead rely on informal networks when hiring. In most countries an ethical board connected to universities will settle whether a particular project will be accepted on ethical grounds.

4. Summary
This article focused primarily on the strengths and limitations of the correspondence testing methodology for studying discrimination in hiring. Although this method cannot address all relevant aspects of labor market discrimination, it can provide strong and direct measures of discrimination when hiring. An important advantage of this testing method is its close connection with laboratory-like conditions, enabling a high degree of control over the analysis and putting the behavior of recruiters in focus. Even so there are certain issues that potentially flaws the results coming out of these experiments. For instance, since the recruiters participating in the experiment are not randomly selected the observed level of discrimination may say nothing about how important discrimination is for explaining differences in employment rates across the same groups. It could be that the group being found to be discriminated against in the correspondence testing experiment is very small compared to the number of discriminating employers and unemployed workers can simply sort themselves away from those firms when searching for jobs. It was also discussed to what extent the results would say something about what caused employers to discriminate, on its external validity and the ethics involved. Most of these concerns could be severely lessened if the ideal experimental design is carefully implemented.

In summary, the correspondence testing method enables a quantitative estimation of the scope of the discrimination in hiring, while the results can affect public opinion and ultimately change employer behavior. The results can also be used as a basis for developing policy initiatives and in the creation of legislation combating discrimination. A potential avenue for future research is to make a cross-country comparison, say, within EU. Cross-country comparison from existing studies are complicated by the fact that different designs are being used and, as the above explanation discusses, this makes interpretation difficult regarding which countries should receive more resources to combat discrimination in hiring. Although some difficulties still exists for launching such a project a carefully designed study would probably come a long way.
References


Table 1. Aggregated results from correspondence testing data.

<table>
<thead>
<tr>
<th></th>
<th>High skill jobs</th>
<th>Medium/low skill jobs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of jobs applied to</strong></td>
<td>1070</td>
<td>482</td>
<td>1552</td>
</tr>
<tr>
<td><strong>Neither invited</strong></td>
<td>62</td>
<td>60</td>
<td>122</td>
</tr>
<tr>
<td><strong>Equal treatment</strong></td>
<td>21</td>
<td>21</td>
<td>42</td>
</tr>
<tr>
<td><strong>Only majority</strong></td>
<td>632</td>
<td>398</td>
<td>1030</td>
</tr>
<tr>
<td><strong>Minority</strong></td>
<td>18</td>
<td>57</td>
<td>239</td>
</tr>
<tr>
<td><strong>Majority</strong></td>
<td>217</td>
<td>217</td>
<td>434</td>
</tr>
</tbody>
</table>

Callback rates:

<table>
<thead>
<tr>
<th>Relative callback rate</th>
<th>Invited minority</th>
<th>Invited majority</th>
<th>Neither</th>
<th>Equal</th>
<th>Only minority</th>
<th>Only majority</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.35</td>
<td>0.26</td>
<td>0.16</td>
<td>0.06</td>
<td>0.6</td>
<td>0.35</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Notes: The table numbers are recalculations from the numbers in Table 1 in Carlsson and Rooth (2007).
Appendix: Example of a job application (translated from Swedish)

Hi,

My name is Erik Johansson and I am 27 years old. I live in Stockholm with my girlfriend Anna. I work as a system designer at Telenor AB in an environment based on win2000/SQL Server. I participate in three different projects and my work involves development, maintenance and everyday problem-solving. Development work is done in ASP, C++ and Visual Basic and we use the development platform .Net and MS SQL. In addition, I have experience in HTML, XML, J2EE and JavaScript.

I enjoy working on development and problem-solving, and I now hope that I will develop further at your company. To my personal characteristics, one could add that I find it easy to work both on my own and in a group. I am a dynamic person who likes challenges. I really like my occupation, which I think is mirrored in the work I do. I have a degree in computer engineering. I graduated with good grades from Stockholm University.

I also like running. It is important for me to keep my body in shape by exercising regularly. Anna and I also like to socialize with our friends during weekends.

I look forward to being invited to an interview and I will then have my certificates and diplomas with me.

Best regards

Erik Johansson/Mohammed Said