DAN-OLOF ROOTH
Linnaeus University, Sweden, and IZA, Germany

I Z A
World of Labor
Evidence-based policy making

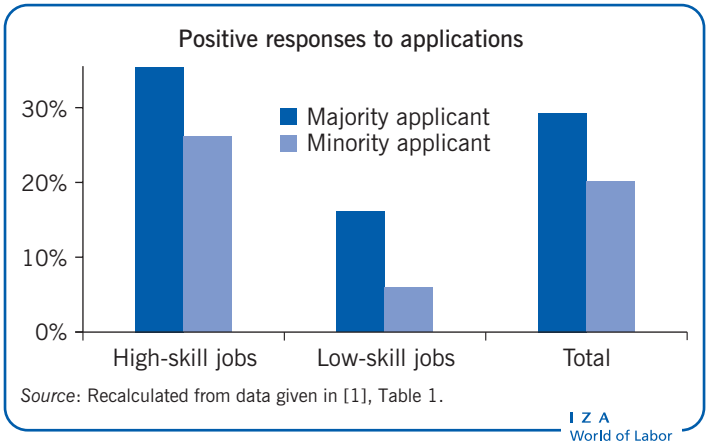
Correspondence testing studies

What can we learn about discrimination in hiring?

Keywords: correspondence testing, taste-based discrimination, statistical discrimination

ELEVATOR PITCH

Anti-discrimination 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. The method involves sending matched pairs of identical job applications to employers posting jobs—the only difference being a characteristic that signals membership to a group.



KEY FINDINGS

Pros

- The correspondence testing (CT) method reveals discriminatory practices at the initial stage of the recruitment process.
- The CT method can be used to test for hiring discrimination based on race or ethnicity, gender, age, sexual orientation.
- The CT method can determine if hiring discrimination varies by occupation and/or region.
- The CT method can increase our knowledge of what characterizes recruiters who discriminate.
- The results can guide policies to prevent discrimination in hiring and to inform employers.

Cons

- The CT method only measures discrimination at the first stage in the hiring process, not in wages or promotions.
- The CT method captures only those firms that use formal search methods, which casts doubts about external validity.
- It is not possible to attribute employer preferences as solely responsible for the result.
- Overuse of the CT method might encourage employers to use alternative search methods for workers.
- The CT method should not be used as sole evidence in legal proceedings.

AUTHOR'S MAIN MESSAGE

The correspondence testing method has been used to identify discrimination in hiring based on ethnicity, gender, age, sexual orientation, and looks in a wide array of countries. Its results are important for guiding anti-discrimination policies and to inform employers about their actions. However, its help in understanding group differences in unemployment is limited, since the behavior of the supply side of the labor market is not investigated.

MOTIVATION

European Union (EU) laws clearly prohibit discrimination in employment based on race or ethnicity, gender, age, disability, and religion or beliefs. However, there are reasons to believe 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.

In Sweden, for example, attitude surveys among the general public (as well as of minority groups) indicate that ethnic discrimination is worst against individuals with a Middle Eastern background. Furthermore, unemployment rates for immigrants born in the Middle East have been found to be several times higher than native Swedes, indicating that ethnic discrimination exists in the recruitment process.

However, discrimination is just one possible explanation for group differences in employment. Another is unobserved differences in productivity characteristics across groups, such as language skills or access to networks. Since researchers can seldom account for all such differences in productivity characteristics between groups, it is difficult to identify empirically the extent of discrimination in employment using a standard regression approach [2], [3]. Another strand of literature makes use of laboratory experiments to identify discriminatory behavior, but this methodology is susceptible to questions of external validity [4], [5].

To circumvent these difficulties, researchers have relied on using field experiments specifically designed to test for discrimination in recruitment. The two most common are correspondence testing (henceforth CT) studies and audits.

The typical CT study sends matched pairs of qualitatively identical job applications to employers that have advertised a job opening. The only difference between the pair is a characteristic that signals membership to a group (such as a name common to a particular ethnic group). The degree of discrimination is quantified by calculating the difference in the number of callbacks for job invitations that members of each group receive.

In an audit study, employers interview carefully-matched pairs of job applicants. However, the audit method has been criticized for not being able to exactly match the pair who attend the job interview on all characteristics other than membership to a group. For instance, in the case of ethnic discrimination, one cannot rule out the possibility that some minority applicants are motivated to prove the existence of discrimination. Accordingly, their actions during the interview may bias the results in favor of discrimination [2], [6].

Advocates of the CT methodology, then, would argue that this method provides the clearest and most convincing evidence of discrimination in hiring. But is this true? The remainder of this paper will provide an overview of the CT methodology and discuss its use in policy-making efforts to eliminate discrimination in hiring practices. Although the focus will be on measuring discrimination in labor markets, the discussion could easily be extended to measuring discrimination in, for example, the housing market.

DISCUSSION OF PROS AND CONS

The first study to use the CT method to detect discriminatory hiring practices took place more than 40 years ago [7]. It measured how likely employers in Birmingham (United Kingdom) were to invite a white (majority) applicant over an Asian or West Indian (minority) applicant to job interviews for white-collar jobs. The test was applied to 32 job vacancies and found that white applicants were more than twice as likely (108%) to be called back as Asian applicants, and 13% more likely than West Indian applicants.

Today, with recruiting more and more likely to be done by email or online, researchers are able to send out thousands of job applications to advertised job openings. For instance, over 13,000 resumes were sent to employers in a study of ethnic discrimination in Canada [8]. This development has made CT studies an increasingly popular method for measuring discrimination in the labor market.

Although most CT studies have used names to signal an applicant's membership to a group (and hence determine whether gender or ethnic discrimination is occurring), recent methodological advances are helping researchers to study discrimination on the basis of age, disability, sexual orientation, and appearance. The empirical design for signaling membership to a group becomes far more challenging in these cases.

For instance, two applicants who are significantly different in age are also likely to have significantly different levels of work experience. Also, employers might find it odd and unusual if applicants signaled explicitly that they are gay or disabled, or if they submitted a photograph of themselves with their application. Hence, this "unusualness" might have its own effect on the callback rate, making it difficult to determine whether the type of discrimination the researcher is trying to detect is actually taking place.

CT studies have been used in a large number of countries and across many different demographic groups. They have found evidence of hiring discrimination against various ethnic groups and women in the United States (US), Canada, and many EU countries. They have also uncovered evidence of discriminatory practices based on age and disability as well as appearance and sexual orientation. In the US and the UK, courts allow parties to file legal claims of discrimination based on the results of CT studies. CT methodology has also been used to measure discrimination in other kinds of markets, such as the housing market.

Interpretation of results from a CT experiment

The interpretation issues involved when reading the results of a CT experiment are illustrated in the following CT study which was conducted in Sweden [1]. The experiment is a typical correspondence study in the sense that it sent matched pairs of (qualitatively) identical applications to the employers. The only difference between the pairs was a signal of membership to a group, which in this case was the name of the applicant (Erik Johansson or Mohammed Said).

An important part of the preparation of the job applications is the choice of observable productivity-related characteristics on which to standardize the applications. The objective of any CT study is to include the job-specific productivity characteristics most important to hiring. What those characteristics are may vary from country to country

Figure 1. Aggregated results from correspondence testing data

						Callback rates		
	Number of jobs applied to	Neither invited	Equal treatment	Only majority invited	Only minority invited	Majority	Minority	Relative callback rate
High-skill jobs	1,070	632	218	160	60	0.35	0.26	1.35
Medium-/low-skill jobs	482	398	21	57	6	0.16	0.06	2.67
Total	1,552	1,030	239	217	66	0.29	0.20	1.50

Notes: The table numbers are recalculations from the numbers in Table 1 in Carlsson, M., and D. Rooth. "Evidence of ethnic discrimination in the Swedish labor market using experimental data." *Labour Economics* 14:4 (2007): 716–729 [1].

I Z A
World of Labor

and from occupation to occupation. Another important part of the experiment's design is the choice of sample size: the number of jobs to apply to. Power calculations are used to determine the minimum sample size required for detecting statistically significant levels of discrimination.

The last row of Figure 1 provides the aggregate results of the Swedish experiment. The two applications were sent in response to 1,552 different job openings. In 1,030 cases, neither applicant was called for an interview. In the remaining 522 cases, at least one of the two applicants was invited to interview, while both applicants were called in 239 cases. In 217 cases, only the majority applicant was invited to interview, compared to just 66 cases in which the minority applicant alone was called. Thus, the callback rates for the majority and minority applicants were 29% and 20%, respectively. In other words, the majority applicant was almost 50% more likely to receive an invitation to interview than the minority applicant.

The results of the experiment suggest that the difference in the callback rate was due to firms or recruiters using membership to a group as a decision variable in the selection process. But how can we interpret this result? And how can we use it outside of the experiment to make suggestions for policy?

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 CT study arises from preference- or taste-based discrimination or from statistical discrimination.

CT studies ideally attempt to measure employer preferences and tastes for hiring majority over minority job applicants by controlling for the most important productivity-related characteristics. However, unless the CT study includes all of the

Taste-based discrimination and statistical discrimination in hiring

Almost all economic analyses of discrimination in hiring have focused on taste-based and/or statistical models of discrimination. Taste- or preference-based discrimination refers to the situation in which employers have a preference for not employing minority workers, while statistical discrimination arises when employers have imperfect information about individuals' productivity and therefore use estimates of group productivity when hiring. If the group productivity of minority workers is lower on average than that of majority workers, employers will statistically discriminate in favor of majority job applicants when hiring.

important characteristics in the hiring process (which differ on average across groups), the CT method cannot separately identify the mechanisms that drive discriminatory treatment [6]. Hence, although a carefully designed CT study includes many (but not all) important productivity characteristics, there are uncertainties regarding whether differences in callback rates should be interpreted solely as arising from taste-based discrimination.

It can be argued that the most policy-relevant issue is whether discriminatory practice or treatment is occurring in hiring at all, not how the researcher might classify it. Obviously, it is illegal to discriminate based on tastes and preferences against minority workers. Recruiters are also not allowed to discriminate by using or making assumptions about the individual's qualifications based on supposed group differences in productivity. Any role that such characteristics have in the hiring process for the CT experiment also falls under the legal definition of discrimination. Hence, the group difference in callback rates in a CT study can be interpreted as capturing the combined effects of taste-based discrimination and statistical discrimination.

Not being able to separate out these alternative explanations is certainly a drawback if one wants to decide upon policy measures to prevent discrimination in hiring from happening. The design of the CT experiment is important in this respect, since the richer the set of applicant characteristics, the less likely it becomes that statistical discrimination plays much of a role in group differences in hiring.

External validity: Choice of firms, occupations, and geographical areas

Most CT experiments respond to job advertisements posted by firms in newspapers or online. Unfortunately, these firms are unlikely to represent a random sample of all firms in the market. Therefore, the credibility of an experiment relies on the extent to which the researcher is able to provide information—about the firms and the channels they use to search for workers—that indicates that the firms are indeed representative of the general labor market. For instance, it could be the case that only less discriminatory firms use very public channels, such as want ads in newspapers, which could lead the researcher to understate the likelihood of discrimination in that market.

The relevance of the CT experiment in terms of its external validity also increases if more occupations important to the labor market are included. The objective when

choosing which occupations to include in an experiment is to get a representative picture of the overall labor market, while at the same time designing a study that is feasible to implement in practice.

To get a representative picture of the labor market, one would like an experiment to include a variation of occupations, since there could be important differences in discrimination depending on the skill level of the job. Ideally, the experimenter can 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 several occupations in order to get a picture of how discrimination varies by occupation.

The first two rows in Figure 1 show the separate results for high- and medium-/low-skill occupations in the Swedish CT study [1]. If this study had included only medium-/low-skill occupations, the conclusion would have been that discrimination is much more severe than if only the high-skill occupations were included.

A related issue is that most CT experiments are restricted to a specific geographic area. This may limit a study's ability to ascertain whether there is a geographical variation in the degree of discrimination. In addition, the results of the experiment are also period specific and could alter as a result of macroeconomic changes, for example, when the labor market tightens.

Although no CT experiment has been able to arrive at a completely random sample of employers, some studies have collected just about all jobs posted within a year in certain occupations and in specific geographical areas. Hence, it is probably fair to say that if minority workers use these channels for their job search, they would encounter the level of discrimination estimated by these studies. Still, it should be emphasized that comparisons of the level of discrimination across studies are complicated by differences in study design, for example, the choice of occupations.

Is this the discrimination being observed in the market?

Even if all firms in the labor market could be included in a CT experiment, the measured level of discrimination might not say much about the probability of whether a minority candidate can actually find a job. It could be the case that many employers have a preference against hiring minority workers, but that these employers are never approached by minority workers and therefore do not have an effect on the probability of their finding jobs.

One researcher asserts, "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" [6].

Despite how different groups sort in the labor market, politicians might still be interested in knowing whether discrimination exists toward a certain group in a particular part of the labor market before proposing a policy that will affect group ratios in employment. For instance, many countries advocate more balanced gender

ratios in the labor market. Policymakers might therefore be interested in knowing whether barriers to such a policy exist, and in what parts of the labor market, before implementing it.

An additional identifying issue

There exists another type of identification problem, one related to group differences in the variance of unobserved—or left out—productivity characteristics. CT studies can obtain biased estimates of discrimination (in any direction) if employers evaluate applications according to some threshold level of productivity [6]. If the variance of unobserved productivity characteristics differs across groups, even though there is no difference in the mean of those same characteristics, this implies that one group has a higher probability to reach over, or fall under, the threshold used for hiring. In fact, in such a scenario, a standard correspondence study could find discrimination when it does not exist or find no discrimination when it does exist. How large this bias might be depends on the design of the correspondence study—specifically, what level of productivity is assigned to applications by the experimenter relative to the threshold potentially used by employers.

This issue has largely been ignored in the empirical literature on CT experiments until the recent appearance of a methodology which may reveal to what extent this criticism of the CT method is empirically justified [3].

Benchmarking discrimination

A recent advance in the empirical design of CT experiments makes it possible, for example, to benchmark the level of discrimination found to the estimated return from job experience. This methodology requires that the researcher not only randomly varies the characteristic that signals membership to a group but also job experience or other relevant labor market skills. Hence, in addition to signals of membership to a group, the applications have different numbers of years of work experience randomly attached to them. This benchmarking makes it possible to ask questions such as, "How does gender discrimination in hiring relate to the return from one extra year of job experience?" The drawback of implementing this design is that it requires contacting a larger sample of employers compared to the standard CT study in order to make a statistically significant inference.

LIMITATIONS AND GAPS

Should correspondence testing be used to identify discriminating employers?

CT studies are permitted as evidence in courts of law in both the US and the UK. However, the results from the Swedish experiment indicate that they should not be the sole evidence [1].

The question to be asked is whether the CT method can be used to prove whether a specific firm has truly discriminated. In other words, is the CT experimental procedure useful in determining whether a single firm consciously chose one applicant over the other in a discriminatory manner? In the Swedish study, for example, only the minority

applicant received a callback 4% of the time. Why that happened could have several explanations.

First, it is possible that some employers have a simple preference for hiring the minority over the majority applicant. However, an examination of the data shows that all 66 of the recruiters were of the majority background. So even if this were true, alternative explanations exist. The applications were sent in random order—and it could be that the candidate whose application arrived first received the only callback. However, in some of these cases the minority application was sent last and that candidate still received a callback.

It could also be the case that the employer/recruiter overlooked the applications which arrived early for some unrelated reason. In other words, it is possible that the CT result will incorporate some randomness at the firm level. However, it should also be said that this randomness plays a minor role in determining the average level of discrimination in the CT experiment, since this randomness affects both groups equally. Nevertheless, this discussion suggests that this type of CT data should not be the only piece of evidence presented in legal cases.

Can we learn something about those who discriminate?

Adding information about employers and their workplaces can be useful for guiding future studies and to learn more about what characterizes those who discriminate. However, because firms are not randomly selected in CT studies, it is problematic to state any causal effects of certain recruiter and/or firm attributes on discriminatory practices when hiring. For instance, in the Swedish study, when information on the gender of the person responsible for hiring and the size of the firm is added, we find that discriminatory practice is largely a male phenomenon that occurs in small firms [1]. However, this result might have little to do with these attributes since these characteristics are not randomly varied. Discrimination might, or might not, disappear if all male recruiters were to be replaced by females.

There are also CT studies that attempt to measure the attitudes of recruiters. Arab-Muslim job applicants are significantly less likely to be interviewed when the recruiter responsible for hiring has stronger negative implicit associations toward Arab-Muslim men [9]. This suggests that automatic processes may exert a significant impact on employers' hiring decisions. However, this result is subject to the same skepticism as above since the recruiters, and hence, their implicit associations, are not randomly varied.

Despite the concern that these additions to the experiment do not produce causal effects, the results could still be useful when designing more controlled laboratory experiments to investigate why employers discriminate in their hiring practices.

Discrimination at other stages

CT studies measure discrimination only in the first stage of the hiring process (who gets called in for an interview). They are not able to capture unequal treatment in who actually gets the job, in promotions or in wage growth. Other methods must be used to study those dimensions.

Ethical concerns

In CT experiments, employers are approached by fictitious job applicants who do not want employment. Nor have the employers been asked to participate in the experiment. The discussions on ethics for CT studies therefore revolve around the issue of deception and the absence of informed consent (and to some extent also the costs—time and legal—born by subjects). It could be argued that, "no harm results from labor market field experiments, because individuals are not identified on publication, and inconvenience to employers and genuine applicants is minimized by offers of interview or employment being promptly declined." and additionally, "that there can be no legitimate expectation of privacy in the act of hiring labor, as national governments and international bodies have accepted the onus of ensuring equality of opportunity for all citizens by declaring discrimination in employment unlawful" [10].

There is an opposing view that non-deceptive practices constitute a public good. If researchers extensively used deception, this might change subject behavior and make experiments harder to interpret. For instance, employers might refrain from announcing job vacancies in newspaper want ads and instead rely on informal networks when hiring. In most countries, an ethics board connected to universities settles whether a particular project is ethical or not, that is, whether the benefits of any particular research study outweigh the costs involved.

SUMMARY AND POLICY ADVICE

While the CT method cannot address all relevant aspects of labor market discrimination, it can provide strong and direct measures of discrimination that occur with 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 devalue the results.

While the CT method is highly recommended for detecting discrimination in hiring, any particular study should be first accepted by the relevant ethics review board(s). Its results can affect public opinion and ultimately change employer behavior. The results can also be used as a basis for developing policy initiatives and for creating legislation to combat discrimination.

Acknowledgments

The author thanks Magnus Carlsson, two anonymous referees and the IZA World of Labor editors for many helpful suggestions on earlier drafts.

Competing interests

The IZA World of Labor project is committed to the *IZA Guiding Principles of Research Integrity*. The author declares to have observed these principles.

REFERENCES

Further reading

Harrison, G. W., and J. A. List. "Field experiments." *Journal of Economic Literature* 42:4 (2004): 1009–1055.

Neumark, D. "Ethnic hiring." In: Constant, A. F., and K. F. Zimmermann (eds). *International Handbook on the Economics of Migration*. Cheltenham, UK: Edward Elgar, 2013; pp. 193–213.

Key references

[1] Carlsson, M., and D. Rooth. "Evidence of ethnic discrimination in the Swedish labor market using experimental data." *Labour Economics* 14:4 (2007): 716–729.

[2] Riach, P. A., and J. Rich. "Field experiments of discrimination in the market place." *The Economic Journal* 112:482 (2002): 480–518.

[3] Neumark, D. "Detecting discrimination in audit and correspondence studies." *Journal of Human Resources* 47:4 (2012): 1128–1157.

[4] Fershtman, C., and U. Gneezy. "Discrimination in a segmented society: An experimental approach." *Quarterly Journal of Economics* 116:1 (2001): 351–377.

[5] Levitt, S. D., and J. A. List. "Field experiments in economics: The past, the present, and the future." *European Economic Review* 53:1 (2009): 1–18.

[6] Heckman, J. J. "Detecting discrimination." *Journal of Economic Perspectives* 12:2 (1998): 101–116.

[7] Jowell, R., and P. Prescott-Clarke. "Racial discrimination and white collar workers in Britain." *Race & Class* 11:4 (1970): 397–417.

[8] Oreopoulos, P. "Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand résumés." *American Economic Journal: Public Policy* 3:4 (2011): 148–171.

[9] Rooth, D. "Automatic associations and discrimination in hiring: Real world evidence." *Labour Economics* 17:13 (2010): 523–534.

[10] Riach, P. A., and J. Rich. "Deceptive field experiments of discrimination: Are they ethical?" *Kyklos* 57:3 (2004): 457–470.

The full reference list for this article is available from the IZA World of Labor website (http://wol.iza.org/articles/correspondence-testing-studies).