HAL open science

# Testing (or correspondence study method)

Nicolas Jacquemet

▶ To cite this version:

Nicolas Jacquemet. Testing (or correspondence study method). LIEPP Methods Brief n°6, 2023. hal-04094048

HAL Id: hal-04094048

https://sciencespo.hal.science/hal-04094048v1

Submitted on 10 May 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

SciencesPo
LABORATORY FOR INTERDISCIPLINARY EVALUATION OF PUBLIC POLICIES

# LIEPP METHODS BRIEF n°6

# 2023

*This brief is part of a set of methods briefs published by LIEPP. As such, it is supported by the ANR and the State under the "Investissements d'avenir" programme within the framework of the IdEx Université Paris Cité (ANR-18-IDEX-0001).*

# Testing (or correspondence study method)

Nicolas JACQUEMET (University Paris 1, CES, PSE)

nicolas.jacquemet@univ-paris1.fr

Distributed under a Creative Commons Paternité.
Attribution-NonCommercial-NoDerivatives | 4.0 International License

https://www.sciencespo.fr/liepp/en.html

How to cite this publication:
JACQUEMET, Nicolas, **Testing (or correspondence study method)**,
*LIEPP Methods Brief* n°6, 2023-05

Publication initialement rédigée en français :
JACQUEMET, Nicolas, **Testing (ou méthode d'étude par correspondance)**,
*LIEPP Fiche méthodologique* n°5, 2023-05

## IN A NUTSHELL:

**Correspondence testing is a quantitative method aimed at measuring discrimination. It involves sending fictitious applications in response to real offers (for example, job offers). By providing and objective measure of discriminatory behaviour, this method is very useful, from a prospective point of view, for designing anti-discrimination policies.**

**Keywords:** Quantitative methods, correspondence, discrimination, anti-discrimination policies, experimental applications

## I. How is this method useful for policy evaluation?

Discrimination refers to an unequal treatment on the basis of individual characteristics that should not be relevant to the decision to be made: favouring a male candidate with superior professional skills is not discriminatory; but rejecting a female candidate on the basis of the suspicion that her availability will be less than that of a male candidate with an equivalent profile is indeed discriminatory, because there is no reason to believe that this female candidate matches this stereotype. As such, discrimination is a major source of inequality. Even more than other types of inequality, discrimination is both very costly in economic terms, by depriving the economy of some of its talent; and persistent, because the anticipation of such inequalities of treatment might discourage the discriminated persons and lead them to make choices (of level and branch of education, of career path) which only amplify these initial inequalities.

Despite the importance of the issues at stake, the development of public policies aimed at combating discrimination suffers from a lack of diagnostic elements due to the great difficulty of measuring it. This is the objective of the "correspondence testing" method.

## II. What does this method consist of?

Although this method can be applied to many different sectors of economic activity (housing search, seasonal rentals, master's students' applications) and to many different sources of discrimination (religion, sexual preference, socio-economic background, place of residence, disability), this presentation focuses for simplicity on its application to the measurement of hiring discrimination based on gender and/or origin.

This method is designed to provide a measure of the success of different types of applicants according to their socio-demographic characteristics, while at the same neutralising the effect of the intrinsic quality of the applications. Each of these two aims has its own methodological implications.

*Constitution of fictitious applications*

The success of different types of candidates is observed through the use of artificial applications, sent in response to real job offers circulating on the labour market. The method combines three ingredients: identities, applications and job offers.

The socio-demographic characteristics whose effect is to be measured are conveyed by the identity of the applicant. In order to test both gender discrimination and discrimination affecting applicants from, for example, North-African origin, a list of four fictitious identities (or four different categories of identities) will thus be created: two French-sounding surnames, one associated with a male first name and the other with a female first name, and two surnames that suggest that the person is from North-African immigration, associated with the same variations of the first name. Each of these identities is given a unique telephone number and an email address to contact the applicants. These first and last names and contact information correspond to the **identity block** of the applications.

In order to respond to job offers, these identities are put forward on applications that most often combine a CV and a cover letter. The aim is to construct applications that are as credible as possible and thus allow to distinguish the success of different identities. The process of constructing applications must therefore lead to a quality that is neither too high nor too low in comparison with the real applications that will be received, because any application that leads to an undifferentiated treatment of applicants, whether positive or negative, makes it impossible to identify the characteristics that favour the success of experimental applications.

The construction of the CV requires to fill the training and experience sections with contents that are realistic and compatible with the job for which the application is sent, as well as a section dedicated to extra-professional activities. In order to ensure that these CV elements correspond to the intended occupations, most studies collect real CVs (available, for example, online), then mix the information from several CVs to construct a unique experimental CV and then modify the resulting 'experience', 'education' and 'extra-curricular activities' sections. The content of the CV is completed by a block containing personal information which includes at least the postal address, but may also mention the marital status, the presence of children, the age or the date of birth. The formatting of this information requires choosing as many predefined templates as there are different CVs, which will determine the order of the sections, the font used and the organisation of the different information (many templates in different file formats can be easily found online). Cover letters are constructed in the same way, combining the content of existing cover letters. The gender arrangements (if any, depending on the language) will be adapted according to the gender of the identity on the application (if gender is one of the characteristics tested, it is advisable to choose formulations that allow for as many occurrences of gender arrangements as possible). The combination of a CV and a cover letter is **an application**.

The **job offers** to which these applications will be sent are collected from public information sites (in France, many studies use the Pôle emploi site, but depending on the profession targeted, it is sometimes necessary to use more specialised ones). These offers are filtered to check that they correspond to the predefined inclusion criteria, which primarily concern the occupation and the location of the job, but also, for example, the requirement of specific experience or skills. The remaining vacancies for which it is not possible to send an application according to the predefined modalities (often by e-mail, but also when, for example, the submission of an application requires the completion of an online questionnaire) are systematically excluded. For the remaining job offers, which will be included in the study, all characteristics of the offer (duration, type of contract, salary, etc.) are carefully recorded in order to build up a database to document the observed heterogeneity of job offers.

The **number of experimental applications** to be sent in response to a given job vacancy (which goes hand in hand with the number of different applications that need to be constructed) is a delicate choice. From a statistical point of view, it is very advantageous to be able to compare the success of different applications in response to a given job offer (i.e. "intra-offer" comparisons), as such comparisons will eliminate the effect of all unobserved elements that are specific to the job offer hence improving the statistical accuracy of the measures. Sending several applications associated to a given socio-demographic group also makes it possible to measure more finely the characteristics of the distribution of discrimination across job offers (see the results presented in Kline et al. 2020). While it is therefore desirable to send several applications in response to each vacancy, the maximum number is limited by two factors. On the one hand, the multiplication of applications increases the disruption caused by the survey on the functioning of the labour market and, above all, the risk of detection. This risk can be contained by taking care to allow sufficient time between the sending of two applications, but this delay goes hand in hand with a reduction in the probability of success for the latest applications, all the more so when the occupation attracts a large number of applications. On the other hand, some recent work (Philips, 2019) shows that the portfolio of applications sent in response to a given offer is likely to affect

the relative success of experimental applications. Increasing the number of applications increases the risk of such a bias.

These two factors together imply to be more restrictive with regard to the number of applications sent, the tighter the occupation. The combination of these different factors leads most studies to limit themselves to sending a maximum of four applications in response to each job offer, sent no later than 24 hours after publication. For this purpose, each identity is associated with a single application (a CV and a cover letter), leading to as many unique and distinct experimental applications as the number of applications sent in response to each vacancy.

Measuring the **success** of the experimental applications requires keeping an accurate, time-limited record (ignoring, for example, responses received more than 3 months after they were sent) of employers' communication with applicants by archiving all written correspondence and transcribing the content of telephone messages received. These responses are then classified to distinguish between refusals, non-responses, requests for further information and invitations to an interview (sometimes referred to as 'expressions of interest'). For obvious ethical reasons, it is imperative to decline any expression of interest as quickly as possible, preferably through the same contact channels and following a predefined script (which most often refers to the previous acceptance of a job offer).

*Assembly protocol*

The combination of all these ingredients provides a measure of the success of fictitious applications that differ, among other things, in the socio-demographic group to which the application identity is associated. Of course, such differences in success can also be linked to the content of the application itself, which is all the more likely when the applications are clearly different from one another. One solution might therefore be to ensure that the applications are as close as possible to each other. But apart from the fact that any difference, however small, between applications would lead to the same conclusion, it is particularly difficult to distinguish between insignificant differences and more important differences, because the differences that are relevant are the subjective variations in the quality of applications that are perceived by employers.

The protocol that allows correspondence studies to neutralise the effect of any potentially confounding characteristic of the experimental applications (i.e. whose impact on the success rate would lead to erroneous conclusions about discrimination) is to systematically rotate the association between identities and applications. If, for example, identity *a* appears on application *A* and identity *b* on application *B* in the first mailing, these associations will be reversed in the next mailing (identity *a* now appearing on application *B*) before returning to the initial association in the third mailing, and so on. This rotation does not eliminate the effect of the perceived quality of the application: if application *A* is found to be of better quality, the success of the identity associated with it will be affected accordingly. But rotation ensures that any systematic difference in identity success across all mailings can no longer be attributed to the content of the application itself. From a statistical point of view, any characteristic for which a systematic rotation is organised becomes a source of noise in the measurement of discrimination related to the characteristics of interest, i.e. a source of variation in the success rate between applications belonging to different categories that is not attributable to discrimination. By construction, this noise is independent of the characteristics whose effect the correspondence study seeks to measure and therefore does not affect the ability of the method to measure discrimination. But it does, however, make it more difficult to detect. These consequences of noise in the measurements can be reduced by adapting the statistical analysis accordingly (in the form of offer fixed effects), but such modelling assumes a homogeneous effect of the quality of applications on all employers.

In sum, the correspondence study method is therefore based on three principles: multiplying the number of experimental applications in order to measure the effect of the socio-demographic characteristics by which they are distinguished, ensuring that these applications are as homogeneous as possible in order to reduce the noise that will affect the measurement of their effect, and organising a systematic rotation of the association between socio-demographic profiles and any other characteristic likely to affect their success. These three principles constitute a toolbox that can be applied to many aspects of the functioning of the labour market. For example, one can measure the effect of unemployment spells in the career path by experimentally modifying the "experience" section of applications, of the distance between the place of residence and work by manipulating the applicant's address, or of the family situation by varying the identity block according to, e.g., the presence of children or the marital status.

## III. An example of the use of this method

A recent study carried out jointly by the Institute of Public Policies and ISM Corum under the aegis of DARES is one of the first large-scale studies to provide an overview of inequalities in access to employment according to gender and origin in the French labour market (Dares IPP and ISM Corum, 2021a and 2021b)). These results confirm that ethnic discrimination is both strong and cross-cutting across all the occupations studied, leading to a penalty of around 30% in the chances of receiving a positive response. This study also highlights the lack of discrimination linked to the gender of the applicant, suggesting that, contrary to a persistent received idea, the strong career inequalities that exist on the labour market between men and women cannot be attributed to hiring decisions.

## IV. What are the criteria for judging the quality of the mobilisation of this method?

The level of the callback rate for a given type of application provides little information about the functioning of the labour market. The results of a correspondence study are rather based on comparisons of callback rates between different types of applications. These comparisons will only manage to detect the difference in success of different types of applicants if the callback rate among reference applications is sufficiently high.

The variations in the socio-demographic characteristics of applicants are introduced through their identity, which is assumed to affect employers' perceptions. To ensure this, it is increasingly common in testing studies to first run a preliminary survey in which a sample of respondents is asked to associate a gender and/or origin with each of the identities presented to them. This survey provides an empirical measure of the quality of the perceptions induced by the identities, and can be used to select the identities included in the study by retaining those whose perceptions are most consistent with the desired group. Such a survey can also be an opportunity to collect additional information on the perceived profile of the identities presented: recent work shows that identities convey many stereotypes linked, for example, to social class or area of residence, which may contribute to the observed differences in success of applications from different categories (Gaddis, 2017).

Finally, observed differences in callback rates are subject to the famous criticism known as the 'Heckman critique', according to which differences in perceived skill variance within different population groups would be sufficient to produce systematic differences in average callback rates, and would be misinterpreted as a systematic bias against these population groups. This critique can be addressed if enough differences in quality are implemented across experimental applications: the statistical analysis can then allow for group-specific variances in unobserved heterogeneity (Neumark, 2012).

## V. What are the strengths and limitations of this method compared to others?

Thanks to its design, the correspondence testing method provides a precise and convincing measure of the extent of discriminatory practices and the specific effect of applicants' socio-demographic characteristics on their successful integration into the labour market. As such, its objectifies a phenomenon that is more difficult to reveal though qualitative approaches: these practices are not easily verbalised in a semi-structured interview, for example, because they are illegitimate and sometimes unconscious. Its main advantage is that it guarantees by construction the independence of these characteristics from all the other elements embedded into the application. The main alternative is to use survey data to study differentials in career paths on the labour market between different categories of the population. But such studies require strong, and often not very credible, statistical assumptions that are needed to neutralise the effect of differences in education or career paths that distinguish these groups and contribute to the observed differences in labour market success.

The scope of the results produced by this method is nevertheless limited by two important factors.

The first is that the success of applications is measured in terms of whether or not they are invited to a job interview. This, however, is a rather imperfect reflection of the final outcome of the recruitment process: the existence of discrimination at this stage of the process only predicts discrimination at the actual hiring stage if all invited candidates are treated equally. If, on the contrary, additional discrimination occurs during the interview process, the measures of discrimination provided by this method underestimate the phenomenon. If, on the other hand, it turns out that the populations discriminated against in the selection of applications are favoured in exactly the opposite proportions when the final candidate is chosen, then these measures distort the reality of discrimination. Audit methods, which consist of using actors playing the role of experimental but real candidates, make it possible to overcome this limitation, but they have the disadvantage of involving a very broad set of factors (physical appearance, voice) that are likely to influence the recruitment process but cannot be distinguished from the socio-demographic characteristics that are apparent.

The second limitation is common to any empirical study but is particularly acute in the case of testings: as discussed above, the more homogeneous the applications, the more accurate the measurements. There are also practical reasons, linked to the fact that the number and specificity of fictitious applications increase with the diversity (geographical or in terms of occupations) of job offers. As a result, testing studies are often limited in scope, and their results can only be conditional on the scope of the study in terms of type of job, sector of activity, geographical area, age range of applicants, etc. The generalisation to the entire labour market of the results observed in this type of study is therefore based on the assumption that the scope chosen does not present any specificities in terms of propensity to discriminate (recruiters' preferences, degree of competition in recruitment, etc.) or, more convincingly, on the accumulation of concordant studies on different spheres of the labour market.

## Some bibliographical references to go further

Adamovic, Mladen. 2020. « Analyzing Discrimination in Recruitment: A Guide and Best Practices for Resume Studies ». *International Journal of Selection and Assessment* 28, no 4 (2020): 445-64.

Adida, Claire. and Laitin, David. and Valfort, Marie-Anne. 2010. Identifying barriers to Muslim integration in France. *Proceedings of the National Academy of Sciences* 107, 22384–22390.

Dares IPP and ISM Corum, 2021a, « Discrimination à l'embauche selon le sexe : les enseignements d'un testing de grande ampleur », *Dares Analyses* n°26/ *Note IPP* n°67

Dares IPP and ISM Corum, 2021b, « Discrimination à l'embauche des personnes d'origine supposée maghrébine: quels enseignements d'une grande étude par testing? *Note IPP* n°76 / *Dares Analyses* n°67

Edo Anthony. and Jacquemet, Nicolas. 2013. La discrimination à l'embauche, Sur le marché du travail Français. *Opuscule du CEPREMAP* n°31, Editions rue d'Ulm

Kline, Patrick. and Walters, Christopher. 2020. « Reasonable doubt: Experimental detection of job-level employment discrimination ». *Econometrica* 89, no 2 (2020): 765-92.

du Parquet, Loïc. and Petit, Pascale. 2019. « Discrimination à l'embauche : retour sur deux décennies de testings en France ». *Revue française d'economie* Vol. XXXIV, no 1 (2019): 91-132.

Gaddis, Michael. 2017. « How Black are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies ». *Sociological Science* 4 (2017): 469-89. https://doi.org/10.15195/v4.a19.

Neumark, David. 2012. « Detecting Discrimination in Audit and Correspondence Studies ». *Journal of Human Resources* 47, no 4 (2012): 1128-57.

Phillips, David. 2019. Do Comparisons of Fictional Applicants Measure Discrimination When Search Externalities Are Present? Evidence from Existing Experiments. *Economic Journal* 129, 2240–2264.

Fougère, Denis. and Rathelot, Roland. and Aeberhardt, Romain. 2011. « Commentaire: Les méthodes de testing permettent-elles d'identifier et de mesurer l'ampleur des discriminations ? » *Economie et Statistique* 447, no 1 (2011): 97-101.