Verhaeghe, Pieter-Paul; Van der Bracht, Koen

Working Paper

How many correspondence tests are enough to detect discrimination among single agents? A longitudinal study on the Belgian real estate market

GLO Discussion Paper, No. 678

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Verhaeghe, Pieter-Paul; Van der Bracht, Koen (2020) : How many correspondence tests are enough to detect discrimination among single agents? A longitudinal study on the Belgian real estate market, GLO Discussion Paper, No. 678, Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/224764

How many correspondence tests are enough to detect discrimination among single agents? A longitudinal study on the Belgian real estate market

Pieter-Paul Verhaeghe* & Koen Van der Bracht

Department of Sociology, Vrije Universiteit Brussel, Belgium

Abstract

Correspondence tests have been used by scholars and civil rights organizations to measure ethnic discrimination. In contrast to research testing covering a whole market through many discrimination tests, litigation testing typically targets a single agent, which can only be tested through a very low number of tests per agent. This low number of tests poses serious methodological challenges to disentangle systematic discrimination from random treatment. This study examines from a purely statistical point of view how many discrimination tests per single agent are needed to convincingly proof discrimination. We collected unique longitudinal data about 114 real estate agents, which were tested through 10 repeated pairwise matched correspondence tests. It appears that 10 or more tests are needed per realtor to detect discrimination with a high degree of certainty. The required number of tests per agent depends on the pattern of discrimination among the agent under study, the expected non-response rate and the desired degree of certainty.

Key words: discrimination; discrimination tests; mid-p-value; longitudinal study; housing market; enforcement testing

JEL Codes: J70, J78, R38

* Corresponding author is Prof. Pieter-Paul Verhaeghe: Pieter-Paul.Verhaeghe@vub.be

1. Introduction

Ethnic discrimination refers to the unequal treatment of people because of their ethnic origin. In most Western countries ethnic discrimination is illegal. Nonetheless, many studies have documented continuing discrimination of minority groups in labor, housing and consumer markets (Rich 2014; Zschirnt and Ruedin 2016; Baert 2018; Neumark 2018; Flage 2018). Anti-discrimination laws have not been able to stop discrimination, but have resulted in a changed manner of discrimination. People rarely discriminate bluntly, but try to exclude certain ethnic minorities by not responding to their job or housing applications, telling minority candidates that the job opening or rental property is no longer available, or offering them less assistance.

Therefore, scholars and policy makers have come up with ways to circumvent this covert discrimination in order to reliably measure and tackle discrimination. They have designed discrimination tests in which pairs of candidates for jobs, accommodation or services are matched in such a way that they do not differ on all relevant characteristics, except on their ethnic origin. Afterwards, the answers to both candidates are compared and an unequal treatment between candidates is assumed to be due to discrimination. These tests are successfully used by scholars to examine discrimination in labor, housing and consumer markets (Rich 2014; Zschirnt and Ruedin 2016; Baert 2018; Neumark 2018; Flage 2018). The use of this methodology is not limited to measure the extent of discrimination for research purposes, but has, for instance, also been used by fair housing organizations or fair employment councils to enforce anti-discrimination laws (often complaint-driven) in the case of a single lessor or employer (Bendick 1998; Rorive 2009; Temkin, McCracken and Liban 2011).

Although research and enforcement testing are based on the same core methodology, they differ in fundamental ways (Fix and Turner 1998; Ross and Yinger 2006; Pager and Western 2012). Whereas research testing is based on a large number of discrimination tests covering the whole market, enforcement testing typically targets a single agent through only a small number of tests.[1] It is both unfeasible and unethical to subject single agents to a similar amount of tests as a whole market. It is unfeasible because an agent only has a limited number of available spaces. It is unethical because the

[1] With agents, we mean firms, organizations and individuals responsible for selection, such as a landlord or a recruiter.

economic burden of extensive testing would be too high to bear for an agent. As a consequence, enforcement tests typically comprise only a very low number of tests.

This small number of tests per agent poses a serious challenge to disentangle discrimination from a random unequal treatment of candidates. There are several situations possible in which agents treat similar candidates unequally without discriminating (Heckman and Siegelman 1993; Ross and Turner 2005). For example, when a lessor receives too many solicitations to process, he can decide to invite only the first ten candidates. If these ten invited candidates contain, simply by chance, more majority candidates and the next group of uninvited candidates more minority candidates, minorities would be treated unequally without being discriminated. Another example is when an employer or landlord forget to answer a few applications and these are, accidently, more likely to come from minority applicants. Finally, an advertised house may be rented out during the interval between two testers. Of course, the reverse situation may also happen, in which minorities are treated, by hazard, more favorably than natives. In the case of a large number of discrimination tests, these random unequal treatments rule each other out statistically, but with a small number of tests this is not the case. Therefore, distinguishing discrimination from random unequal treatment by single agents is an important methodological issue to consider for policy makers and social scientists.

Policy makers who want to act against discriminatory behavior may be inclined to use discrimination tests as a way to prove discrimination. To be successful in court, the claimant has to bring clear and convincing evidence of discrimination to meet the requisite standard of proof. In many countries claimants have to convince the court of a 'presumption of discrimination'. In contrast to the US experience, many European courts are rather reluctant to accept discrimination tests as convincing evidence of a presumption of discrimination (Rorive, 2009). The difficulty to disentangle systematic discrimination from a random unequal treatment by single agents, due to the small number of tests, may be a reason for this reluctance.

For social scientists, determining discrimination per agent may give insights into the pattern of discrimination. It is still unclear whether the discrimination levels in many countries are the result of discrimination against minorities throughout all agents under study, or only among a small group of agents. What causes some agents to be more prone to show discriminatory behavior than others? Is discrimination by realtors or employment agencies instigated by only some clients with specific preferences or do they discriminate for all advertisements? Therefore, developing a consistent

methodology to measure discrimination among single agents may open up new research possibilities for scholars interested in patterns of ethnic discrimination.

The objective of this study is, hence, to develop a method of discrimination tests for single agents and examine how many discrimination tests per agent are needed to convincingly determine the occurrence of discrimination. We consider the significance levels of discrimination rates as indicators of the strength of proof along a continuum: systematic discrimination is more presumed with higher p-values and, vice versa, more certain with lower p-values. Afterwards, we simulate how many discrimination tests per agent are necessary to detect discriminating agents. For these purposes, we performed a unique longitudinal study among 114 real estate agents in Belgium. We monitored their potential discriminatory behavior during ten waves, conducting 1140 correspondence tests. To the best of our knowledge, this is the first longitudinal correspondence test study that examines an actor during 10 successive waves.

2. The Method of Discrimination Tests

At its core, the design of discrimination tests is rather simple: two candidates apply for a house, job or service. The candidates are similar on all relevant characteristics, except for the trait under scrutiny. One candidate originates from the test group (ethnic minority), the other from the control group (ethnic majority). Afterwards, scholars examine whether both candidates were treated equally. Unequal treatment is assumed to be due to discrimination.

When two pairwise matched candidates apply for a vacant job or house, there are four possible outcomes: both candidates are invited for a job interview or to visit the dwelling (n_{11}), only the control (ethnic majority) person is invited (n_{21}), only the test (ethnic minority) person is invited (n_{12}), or neither are invited (n_{22}). This results in a two-by-two contingency table, as demonstrated in Table 1.

TABLE 1 ABOUT HERE

The most straightforward measure of discrimination would be the proportion of all tests in which the control candidate is favored over the test candidate (n_{21}). This gross rate overstates discrimination,

however, because at least part of the majority-favored treatment may be attributed to random factors, as explained in the introduction. It is assumed that no random treatment exists and that any unequal treatment is a form of discrimination.

Therefore, the net rate of discrimination subtracts the minority-favored treatment (n_{12}) from majority-favored treatment (n_{21}). The assumptions, here, are that all cases of minority-favored treatment are due to random factors and that random majority-favored treatment occurs just as frequently as random minority-favored treatment (Ross and Turner 2005). Minority-favored treatment may, however, be systematic, for instance due to minority landlords preferring to rent to people of their own ethnic group as well as realtors specializing in minority clients or steering majority candidates away from housing in minority neighborhoods. This means that net rates of discrimination understate systematic discrimination (Ondrich, Ross, and Yinger 2000; Ross and Turner 2005). This argument is, however, less convincing in the case of single agent testing, since it is unlikely that a single agent is systematically discriminating both against and in favor of ethnic minorities. As a consequence, most 'reverse' discrimination of a single agent will be due to 'random' factors. Nevertheless, most studies consider gross and net rates of discrimination as upper and lower bounds of discrimination.

In addition, there is some debate about the situation when neither candidate is invited (n_{22}). This outcome can be considered as equal treatment or as non-response (Riach and Rich 2002). The latter approach generates in general higher rates of discrimination than the former. In line with the International Labour Organization (Bovenkerk 1992), most studies treat this outcome as non-response (Riach and Rich 2002), since there is no information at all about whether there are discriminatory intentions or not. Following these arguments, the gross and net rates of discrimination are calculated as follows:

$$Gross\ rate\ of\ discrimination = \frac{n_{21}}{n_{11} + n_{12} + n_{21}}$$

$$Net\ rate\ of\ discrimination\ = \frac{n_{21} - n_{12}}{n_{11} +\ n_{12} + n_{21}}$$

3. Testing the Significance of Discrimination Rates

For both gross and net rates of discrimination, we need to ascertain that the rates per agent we found in the sample of tested vacancies are statistically discernable from zero in the population of all vacancies of that single agent. For the gross rate, the null hypothesis of no discrimination against ethnic minorities in the population is $H_0: \pi_{21} = 0$. The null hypothesis states that there are no vacancies of a particular agent where the ethnic majority candidate would receive an invitation and the minority candidate not. Although an appropriate test statistic can be calculated to test this null hypothesis, testing this hypothesis is meaningless (Turner et al. 2002). As stated above, the gross rate of discrimination assumes that no random treatment exists and that any unequal treatment is a form of discrimination.

For the net discrimination rate, the null hypothesis of no discrimination against ethnic minorities in the population is $H_0: \pi_{21} = \pi_{12}$, or $H_0: \pi_{21} - \pi_{12} = 0$. The null hypothesis states that unequal treatment of minority and majority candidates is equal in the population of all advertisements of a particular agent. There may be unequal treatment between both candidates, but both candidates are subjected to the same proportion of unequal treatment, therefore singling each other out over different vacancies per single agent. To test this hypothesis, we need a small sample test designed for matched-pair data. The former is a crucial aspect, since we will be working with very small sample sizes. To test a statistically significant difference between dichotomous matched-pair data, a McNemar test is usually applied. This test-statistic follows a chi-squared distribution with df=1. This is an asymptotic test, however, typically used in large sample situations. In small samples, the required asymptotics do not hold, leading to violations of the nominal significance level (Fagerland, Lydersen, and Laake 2013).

A possible solution is using a McNemar mid-p test for matched-pair dichotomous data (Lancaster 1961). The mid-p-value is obtained by including half the probability of finding the observed point in each tail. Therefore, a two-sided McNemar mid-p-value can be calculated as follows:

$$\left(\sum_{x_{12}=0}^{\min(n_{12}, n_{21})} \binom{n}{x_{12}} \left(\frac{1}{2}\right)^n \right) - \binom{n}{\min(n_{12}, n_{21})} \left(\frac{1}{2}\right)^n \quad (1)$$

Mid-p tests have been termed quasi-exact (Hirji et al. 2011). The test is based on an exact distribution, but it is not guaranteed that the nominal significance level is not exceeded. In a simulation study, Fagerland and his colleagues (2013), compare the performance of the McNemar mid-p-value to that

of the asymptotic McNemar test and the McNemar exact test. In none of the 9595 simulations the nominal significance level was exceeded by the mid-p test. Therefore, the McNemar mid-p test is an excellent test to determine the occurrence of discrimination: it is not over conservative for small samples such as the McNemar exact test and it is not likely to violate the nominal significance level for small samples, such as the asymptotic McNemar test.

The p-value of the net-discrimination rate, calculated by the mid-p test, is thus a function of two parameters: n_{12} and n_{21}. The p-values for all possible values of n_{12} and n_{21} under the condition of $n_{12} + n_{21} \leq 10$ are presented in table 2. If a scholar or policy maker takes the traditional 0.05 significance level as the critical value, H_0 of no discrimination would only be rejected when n_{12}= 0 and $n_{21} > 4$ or when n_{12}= 1 and $n_{21} > 6$. If they use, however, a less conservative 0.10 significance level as the critical value, the null-hypothesis would only be rejected when n_{12}= 0 and $n_{21} > 3$, when n_{12}= 1 and $n_{21} > 5$, or when n_{12}= 2 and $n_{21} = 8$. The choice of a particular significance level is off course an important issue to consider in determining the necessary number of discrimination tests to proof discrimination.

TABLE 2 ABOUT HERE

A significance level of 0.05 or lower is usually used in social sciences as a rule of thumb. If studies yield net discrimination rates below this significance level, most scholars would agree to speak about discrimination, especially if the significance level is found among small samples of maximum ten tested vacancies. There are, however, two reasons why less conservative significance levels may be used. First, many studies have documented substantial discrimination against ethnic minorities (Rich 2014; Zschirnt and Ruedin 2016; Flage 2018). Therefore, a one-sided test, or alternatively an adoption to the 0.10 significance level could be defended. Second, in the case of litigation testing, we are only looking for 'presumptions' of discrimination. Whereas a 0.05 significance level provides more convincing evidence of discrimination, this threshold may be too conservative to only 'presume' discrimination. In this case, higher threshold values to evaluate the p-value may be used. Therefore, these significance levels of 0.05 and 0.10 should be considered as possible thresholds in a continuum of the strength of evidence.

4. Data and Methods

We performed a longitudinal study among real estate agents in Belgium. In March 2015, we randomly selected 320 realtors with advertisements for private rental houses in the North of Belgium from a major real estate advertising website.[2] In ten successive waves between April and October 2015, we randomly selected one advertised dwelling per realtor per wave. In each wave, we conducted a pairwise matched correspondence test per realtor for the selected advertised dwelling of that realtor. However, some realtors with a low number of properties did not have a property available at each wave of data collection, meaning that we had to remove these realtors from the data. In the final longitudinal sample, we retain 114 of the original 320 realtors (35.6%).[3] This means that we have 10 pairwise matched correspondence tests of ethnic discrimination for each of these 114 realtors.

Realtors were contacted via e-mail by a matched pair of fictitious, male candidates. The test candidate was someone with an Arabic-sounding name (Arabic-speaking people are the largest non-European minority group in Belgium), the control candidate had a regular Flemish-sounding name (Flemish people are the ethnic majority in the North of Belgium).[4] The use of names to signal the ethnic origin of a candidate is a common practice in discrimination research (Carpusor and Loges 2006; Verhaeghe et al. 2017). Each candidate asked whether the dwelling was still available and whether he could inspect the property. Both e-mails were semi-identical and were sent at almost the same time to the realtors.

To examine which agents are systematically discriminating, we calculate gross and net discrimination rates for each realtor with 10 correspondence tests. Significance levels of net rates are calculated using mid-p test statistics for pairwise matched dichotomous data (formula 1). Once we have defined through 10 correspondence tests which agents are significantly discriminating according to a certain significance level, we can afterwards determine the minimum number of discrimination tests necessary to detect these agents. For this purpose, we perform simulations of what would happen if we would have sampled a lower number of correspondence tests. For each agent, we draw observations from the available data and recalculate the net discrimination rate for situations whereby

[2] The website Immoweb.be is the major real estate advertising website in Belgium with over 150,000 real estate advertisements.

[3] Attrition analyses reveal that the net discrimination rate is slightly lower among the smaller real estate agents, while there is no difference in the gross rate.

[4] Examples of Arabic sounding names are Fahim Amhali, Houssam Idrissi or Mustafa Atalik. Examples of Flemish sounding names are Erik Debruyne, Pieter Coppens, or Tom Vermeulen.

n is equal to 9, 8, 7, 6, 5 and 4. Scenarios whereby $n < 4$ are not calculated, since p-values always exceed 0.10 when $n < 4$. In each scenario, there are a number of unique combinations. This results in 847 simulations. For each simulated combination, we calculate the net-discrimination rate and corresponding mid-p value. Afterwards, we measure in what proportion of tests discrimination among real estate agents would go undetected if we lower the number of correspondence tests. These simulations are only performed for those real estate agents that have a net discrimination rate below the threshold p-values of 0.10 in the data of 10 correspondence tests.

5. Results

The results of the correspondence tests among 114 realtors are depicted in table 3. The table is sorted on p-values and discrimination rates, with the realtors with lowest p-values at the top, reflecting the continuum of the strength of proof of discrimination. Because of space limitations, we only show the first 15 – most discriminating – realtors, but the full table is available as supplementary material online. Among the whole sample of 114 agents, we see that there is substantial discrimination against ethnic minorities in Belgium. The total gross rate of discrimination is 32.5% and the net rate 22.5% (p<0,001). This finding is in line with other Belgian studies on housing market discrimination (Heylen & Van den Broeck 2016; Van der Bracht et al. 2015; Verhaeghe et al. 2017).

TABLE 3 ABOUT HERE

If we look at the results per single agent, we see that realtors 1 to 8 have mid-p-values below 0.05 and realtors 1 to 13 have a mid-p-value below 0.10. For each of these 13 realtors, there is less than 10% probability that we would find discrimination in these 10 correspondence tests while there is no discrimination in the full population of dwellings these realtors offer. There is good reason to accept a p-value of 0.10, because among realtors 9 to 13 the aggregated discrimination rates are still 52.6%.[5]

These results suggest that the problem of ethnic discrimination may be caused by a limited number of realtors with very high rates of discrimination, and not by the behavior of the whole populations of

[5] The gross rate is equal to the net, since n_{12} is zero.

real estate agents. Among the 13 realtors with high certainty of discrimination, aggregated gross and net rates are 69.6% ($p<0.001$), while the aggregated gross rate among the remaining 101 realtors is 24.7% and the aggregated net rate is 12.7% ($p<0.001$). Although ethnic discrimination is hence significantly lower if the realtors who discriminate with more certainty are removed from the sample, discrimination does not disappear altogether. Therefore, ten correspondence tests may be enough to single out those realtors that discriminate fiercely, but more may be needed to eradicate discrimination altogether.

The question is then: could we reduce the number of correspondence tests further and still detect discrimination among those realtors for whom discrimination is relatively certain? To answer this question, we look at the results of the 847 simulations in table 4. The situations whereby significant discrimination would be determined in 100% of the cases are printed in bold. As we can see from table 4, conducting 9 correspondence tests instead of 10 would mean that discrimination among realtors 9 to 13 would only be detected in 60% of the cases. In 40% of all situations where 9 correspondence tests would be conducted, discrimination would go unnoticed for the group of realtors with an aggregated net rate of 52.6%. The percentage of situations whereby discrimination would be detected among realtors 9 to 13 further declines to 33.3% with 8 tests, 16.7% with 7 tests, 7.1% with 6 tests, 2.4% with 5 tests and 0.5% with 4 tests. Similar results are found for realtors 1 to 8. In the case of 9 correspondence tests, discrimination by realtors 6 to 8 would go unnoticed in 50% of the cases if the 0.05 significance level is used. If the 0.10 significance level is applied, all 8 realtors would be found to discriminate significantly in 100% of 9 correspondence tests. This changes quickly, however, in the other simulations. In the case of 8 correspondence tests only the 5 realtors with the highest discrimination rates are detected in 100% of the situations, with 7 correspondence tests only the first 4 realtors, with 6 and 5 only the first realtor and with 4 tests none of the realtors. If we look at the 0.05 significance level, only the first realtor would be detected with 7 and 6 correspondence tests.

TABLE 4 ABOUT HERE

Table 3 also shows how non-response by realtors affects the gross and net discrimination rates, if non-response is excluded from the formula. Compare for instance realtor 14 to realtor 12. Although the latter displays a significant net rate ($p = 0.063$), the net rate of realtor 14 exceeds the 0.10 threshold.

At the same time, the net and gross rate of realtor 14 is higher than that of realtor 12: respectively 100% and 44.4% for both gross and net rates. This is due to the exclusion of non-response: if cell n_{22} would be included in the numerator of the formulas, the gross and net rates would have been 40% for realtor 12 and 33.3% for realtor 14. Including the non-response in the numerator would therefore increase the comparability of the discrimination rates to their corresponding p-values. At the same time, the table shows that non-response can be an important factor in finding significant discrimination among agents. Although non-response is not included in the formula to calculate p-values, we observe that the higher the non-response, the more correspondence tests are needed to retain sufficient valid cases to obtain a significant p-value. This can be easily seen in Table 2: if a realtor only responds to 50% of the paired tests, the possibility of finding significant discrimination is restricted to cases where the minority candidate is discriminated against in 4 or 5 out of 5 cases. Therefore, the number of correspondence tests will need to be higher in situations where non-response is higher.

6. Discussion and conclusion

The aim of this study was to examine how many discrimination tests per single agent are needed to convincingly statistically proof the occurrence of discrimination. It appears that 10 or more correspondence tests are needed per single real estate agent to detect discrimination with a relatively high degree of certainty. From the 114 selected real estate agents, we see that 8 realtors have test statistics with mid-p-values below the threshold of 0.05 and an additional 5 realtors have mid-p-values below the threshold of 0.10. The finding that the overall discrimination rate remains significant after removal of these 13 discriminating real estate agents suggests that even more than 10 correspondence tests are needed to single out all real estate agents that show significant discrimination against ethnic minorities. Lower numbers of correspondence tests per realtor can still be used, but only at the cost of less convincing proof of discrimination. We considered the significance levels of discrimination rates as indicators of the strength of proof along a continuum: systematic discrimination is more presumed with higher p-values and, vice versa, more certain with lower p-values. Because in many countries claimants have to convince courts of a 'presumption of discrimination' (Rorive 2009), lower numbers of correspondence tests per agent, with as consequence higher p-values, may be used. There are, for example, many cases of successful litigation testing in the US based on only a couple of tests per agent. From a strictly legal (and not statistical) point of view, even one test of discrimination could be enough to establish a 'presumption of discrimination'.

An important issue to take into account is the problem of 'false positives' or 'Type I-errors' in the case of multiple significance testing. With p-values we calculate the probability that a certain discrimination rate or any more extreme rate would be found if in reality there is no discrimination. With threshold values of 0.05 and 0.10, we basically agree that in 5% or 10% we could find a similar result if there is no discrimination in reality, i.e. false positives. This is not a problem if only the discrimination rate of one agent is tested. If we repeat the same test multiple times, however, we allow the false positive rate of 5% or 10% to occur multiple times. This means that the probability of observing at least one false positive significant result among the 13 'discriminatory' real estate agents in our sample is rather high: $1-(1-0.05)^{13}=48.7\%$ in the case of a 0.05 threshold and $1-(1-0.10)^{13}=74.6\%$ in the case of a 0.10 threshold.

Therefore, it is recommended to provide – next to the p-values – substantial arguments why false positives would be very unlikely to arise by chance. In this study, there are several of these arguments. Firstly, given that we performed two-sided tests, the expected number of false positives with a p-value

of 0.10 is 5 cases of ethnic *majority* favoring and 5 cases of ethnic *minority* favoring. All 13 real estate agents for whom p-values below the threshold were found, displayed, however, significant discrimination against ethnic minorities, none against ethnic majority candidates. Secondly, the aggregate net and gross discrimination rates we found for the full sample was comparable to previous housing research in Belgium (Heylen & Van den Broeck 2016; Van der Bracht et al. 2015; Verhaeghe et al. 2017), while the aggregate rates of the remaining 101 real estate agents was much lower. It seems unlikely that all thirteen significant discrimination rates are false positives, because these thirteen cases are the main driver of a discrimination rate comparable to previous studies.

At the same time, it is not unthinkable that at least some of these thirteen are false positives. For policy makers, this is not so problematic: as the correspondence tests are only an element to indicate a presumption of discrimination to initiate a complaint, false positive defendants will be able to refute the allegations. For scholars, it suggests that a higher number of tests may be needed to allow the threshold to be corrected for multiple testing purposes. A threshold of 0.05 or 0.10 for only ten cases can, however, already be considered as a fairly conservative approach.

In sum, the required number of tests per agent to convincingly show discrimination depends on three factors: the pattern of discrimination among the agent under study, the expected non-response rate and the chosen threshold p-value. Among agents with less systematic discrimination and higher expected non-response rates, one should use more discrimination tests per agent. Since p-values can be considered as indicators of the strength of proof along a continuum, one should make a trade-off in choosing a particular p-value between the cost and burden of performing many discrimination tests per agent on the one hand and the degree of certainty they wish to have. Once the appropriate number of discrimination tests is determined, upper and lower bounds of discrimination can be calculated using the gross and net rate and the proof can be statistically tested using the McNemar mid-p-value.

A high number of discrimination tests per agent has a few important consequences. Enforcement testing may in some markets only be possible for agents with a high number of available advertisements, such as bigger employment agencies, companies and realtors. This difficulty arose already in this study, since we could only test 114 of the original 320 realtors ten times. In addition, high numbers of discrimination tests per agent requires substantial resources for testing. Therefore, we recommend policy makers to (re)invest in agencies responsible for litigation and research testing (such as more funding for the Department of Housing and Urban Development in the US). The

methodology developed in this study can result in clear and convincing evidence of discrimination, but it takes professional agencies or organizations to put it into place.

References

Baert, S. 2018. “Hiring discrimination: an overview of (almost) all correspondence experiments since 2005.” In: M. Gaddis (ed.). *Audit studies: behind the scenes with theory, method and nuance.* New York: Springer.

Bendick, M. 1998. Adding testing to nation’s portfolio of information on employment discrimination. In: M. Fix & M.A. Turner (eds.), *Report Card on Discrimination in America: the Role of Testing.* Washington, DC: Urban Institute Press.

Bovenkerk F. 1992. *A Manual for International Comparative Research on Discrimination on the Grounds of ‘Race’ and ‘Ethnic origin’.* Geneva: International Labour Organization.

Carpusor, A.G., Loges, W.E. 2006. “Rental discrimination and ethnicity in names.” *Journal of Applied Social Psychology,* 36, 934-952.

Fagerland, M.W., Lydersen, S., Laake, P. 2013. “The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional.” *BMC Medical Research Methodology,* 13, 1-8.

Flage, A. 2018. “Ethnic and gender discrimination in the rental housing market. Evidence from a meta-analysis of correspondence tests, 2006-2017.” *Journal of Housing Economics,* 41, 251-273.

Fix, M., Turner, M.A. 1998. Measuring racial and ethnic discrimination in America. In: M. Fix & M.A. Turner (eds.), *Report card on discrimination in America: the role of testing.* Washington, DC: Urban Institute Press.

Heckman, J.J., Siegelman, P. 1993. The Urban Institute audit studies: Their methods and findings. In: M. Fix, R.J. Struyk (eds.), *Clear and Convincing Evidence: Measurement of Discrimination in America* (pp. 187–258). Washington, DC: Urban Institute Press.

Heylen, K., Van den Broeck, K. 2016. “Discrimination and selection in the Belgian private rental market.” *Housing Studies,* 31, 223-236.

Lancaster, H.O. 1961. “Significance tests in discrete distributions.” *Journal of the American Statistical Association,* 56, 223-234.

Neumark, D. 2018. “Experimental research on labor market discrimination.” *Journal of Economic Literature,* 56, 799-866.

Ondrich, J., Ross, S., Yinger, J. 2000. “How common is housing discrimination? Improving on traditional measures.” *Journal of Urban Economics,* 47, 470-500.

Pager, D., Western, B. 2012. “Identifying discrimination at work: the use of field experiments.” *Journal of Social Issues,* 68, 221-237.

Riach, P.A., Rich, J. 2002. “Field experiments of discrimination in the market place.” *The Economic Journal,* 112, 480-518.

Rich, J. 2014. “What do field experiments of discrimination in markets tell us? A meta-Analysis of studies conducted since 2000.” *IZA Discussion Paper No. 8584.*

Rorive, I. 2009. *Proving Discrimination Cases: the Role of Situation Testing.* Brussels: Migration Policy Group & Centre For Equal Rights.

Ross, S., Turner, M.A. 2005. “Housing discrimination in metropolitan America: explaining changes between 1989 and 2000.” *Social Problems,* 52, 152-180.

Ross, S., Yinger, J. 2006. “Uncovering discrimination: a comparison of the methods used by scholars and civil rights enforcement officials.” *American Law and Economics Review,* 8, 562-614.

Temkin, K., McCracken, T., Liban, V. 2011. *Study of the Fair Housing Initiatives Program.* Washington, DC: US Department of Housing and Urban Development.

Turner, M.A., Ross, S., Glaster, G., Yinger, J. 2002. *Discrimination in Metropolitan Housing Markets: Phase I – National Results from Phase I of the HDS 2000.* Washington, DC: US Department of Housing and Urban Development.

Van der Bracht, K., Coenen, A., Van de Putte, B. 2015. “The not-in-my-property syndrome: the occurrence of ethnic discrimination in the rental housing market in Belgium.” *Journal of Ethnic and Migration Studies,* 41, 158-175.

Verhaeghe, P.P., Coenen, A., Demart, S., Van der Bracht, K., Van de Putte, B. 2017. *Discrimibrux 2017. Discrimination sur le marché locatif privé (agences immobilières) de la Région de Bruxelles-Capitale.* Brussels: Sociology Department – VUB.

Zschirnt, E., Ruedin, D. 2016. “Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests 1990-2015.” *Journal of Ethnic and Migration Studies,* 42, 1115-1134.

Table 1. Unequal treatment contingency table: observed counts (and joint outcome probabilities)

			Control (ethnic majority) person invited?	
			Yes	**No**
Test (ethnic minority) person invited?		**Yes**	$n_{11}\ (p_{11})$	$n_{12}\ (p_{12})$
		No	$n_{21}\ (p_{21})$	$n_{22}\ (p_{22})$

Table 2. mid-p-values for $n_{12} + n_{21} \leq 10$

		n_{12}										
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
n_{21}	**0**	1.000	0.500	0.250	0.125	0.063	0.031	0.016	0.008	0.004	0.002	0.001
	1	0.500	1.000	0.625	0.375	0.219	0.125	0.070	0.039	0.021	0.012	
	2	0.250	0.625	1.000	0.688	0.453	0.289	0.180	0.109	0.065		
	3	0.125	0.375	0.688	1.000	0.727	0.508	0.344	0.227			
	4	0.063	0.219	0.453	0.727	1.000	0.754	0.549				
	5	0.031	0.125	0.289	0.508	0.754	1.000					
	6	0.016	0.070	0.180	0.344	0.549						
	7	0.008	0.039	0.109	0.227							
	8	0.004	0.021	0.065								
	9	0.002	0.012									
	10	0.001										

Table 3. Discrimination rates real estate agents (N = 114) – full table available as supplementary data online

ID	Gross rate	Net rate	Mid-p-value	n_{11}	n_{12}	n_{21}	n_{22}
1	1.000	1.000	0.002	0	0	9	1
2	0.875	0.875	0.008	1	0	7	2
3	0.875	0.875	0.008	1	0	7	2
4	0.875	0.875	0.008	1	0	7	2
5	1.000	1.000	0.016	0	0	6	4
6	0.714	0.714	0.031	2	0	5	3
7	0.556	0.556	0.031	4	0	5	1
8	0.556	0.556	0.031	4	0	5	1
9	0.667	0.667	0.063	2	0	4	4
10	0.667	0.667	0.063	2	0	4	4
11	0.571	0.571	0.063	3	0	4	3
12	0.444	0.444	0.063	5	0	4	1
13	0.400	0.400	0.063	6	0	4	0
14	1.000	1.000	0.125	0	0	3	7
15	0.750	0.750	0.125	1	0	3	6
…	**…**	…	…	…	…	…	…
Total	0.325	0.225	< 0.001	340	59	192	549

Table 4. Simulation results of fewer tests among discriminating realtors

Id	Gross rate	Net rate	p-value	9 tests		8 tests		7 tests		6 tests		5 tests		4 tests	
				p < 0.05	p < 0.10	p < 0.05	p < 0.10	p < 0.05	p < 0.10	p < 0.05	p < 0.10	p < 0.05	p < 0.10	p < 0.05	p < 0.10
1	1.000	1.000	0.002	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.500	**1.000**	0.000	0.600
2	0.875	0.875	0.008	**1.000**	**1.000**	**1.000**	**1.000**	0.708	**1.000**	0.333	0.833	0.083	0.500	0.000	0.167
3	0.875	0.875	0.008	**1.000**	**1.000**	**1.000**	**1.000**	0.708	**1.000**	0.333	0.833	0.083	0.500	0.000	0.167
4	0.875	0.875	0.008	**1.000**	**1.000**	**1.000**	**1.000**	0.708	**1.000**	0.333	0.833	0.083	0.500	0.000	0.167
5	1.000	1.000	0.016	**1.000**	**1.000**	0.667	**1.000**	0.333	0.833	0.119	0.548	0.024	0.262	0.000	0.071
6	0.714	0.714	0.031	0.500	**1.000**	0.222	0.778	0.083	0.500	0.024	0.262	0.004	0.103	0.000	0.024
7	0.556	0.556	0.031	0.500	**1.000**	0.222	0.778	0.083	0.500	0.024	0.262	0.004	0.103	0.000	0.024
8	0.556	0.556	0.031	0.500	**1.000**	0.222	0.778	0.083	0.500	0.024	0.262	0.004	0.103	0.000	0.024
9	0.667	0.667	0.063	0.000	0.600	0.000	0.333	0.000	0.167	0.000	0.071	0.000	0.024	0.000	0.005
10	0.667	0.667	0.063	0.000	0.600	0.000	0.333	0.000	0.167	0.000	0.071	0.000	0.024	0.000	0.005
11	0.571	0.571	0.063	0.000	0.600	0.000	0.333	0.000	0.167	0.000	0.071	0.000	0.024	0.000	0.005
12	0.444	0.444	0.063	0.000	0.600	0.000	0.333	0.000	0.167	0.000	0.071	0.000	0.024	0.000	0.005
13	0.400	0.400	0.063	0.000	0.600	0.000	0.333	0.000	0.167	0.000	0.071	0.000	0.024	0.000	0.005