Research in Social Stratification and Mobility 65 (2020) 100475

Contents lists available at ScienceDirect

Research in Social Stratification and Mobility

journal homepage: www.elsevier.com/locate/rssm

Induced competition in matched correspondence tests: Conceptual and methodological considerations

Edvard N. Larsen

University of Oslo, Department of Sociology and Human Geography, Molkte moes vei 31, Postboks 1069 Blindern, 0371 Oslo, Norway

ARTICLE INFO

Keywords:
Correspondence studies
Discrimination
Experimental methods
Simulation

ABSTRACT

Audit and correspondence studies are established as the dominant empirical strategy for examining the prevalence of hiring discrimination in labor markets. Historically, these studies are most often conducted as pairwise within-subject tests where one of the testers or applicants are exposed to the treatment of interest, while the other serves as a reference. In more recent studies, however, some scholars are moving towards the practice of sending only a single application per vacancy. This design choice is rarely discussed explicitly. Taking experiments on ethnic hiring discrimination as a case, I summarize the strengths, weaknesses and particular considerations associated with both approaches, by aid of a computational model comparing the two. The main argument I present is that the matched and unmatched designs differ in a more substantive way than what has been properly acknowledged in previous studies: Choice of design can potentially influence estimates, as the matched design induces competition affecting the callback probability of the treated candidates when applicant pools are small. I conclude that this is a potential weakness of the matched design, but that each design has multiple distinct advantages. While this study discusses experiments on hiring discrimination specifically, the main argument applies to similar contexts where one has to choose between matched and unmatched designs.

1. Introduction

Due to their ability to convincingly establish causal relationships in ways observational studies cannot, randomized field experiments have taken a central position in research aiming to detect discrimination in various social arenas. Since the early audit studies of the 1960s, where pairs of trained testers were applied to job interviews, the volume and sophistication of these studies have grown substantially (Cherry & Bendick, 2018; Gaddis, 2018). Diverse grounds for discrimination have been subjected to experimental investigation: Race or ethnicity, gender and motherhood, age, sexual orientation, religious beliefs among others. Following the early audit studies using live testers, most research moved on to correspondence studies, where written text is used in place of live actors: The classical example is submitting fictitious job applications for vacancies, exploiting the advantage of having full control over all information presented to the employer (see Jowell & Prescott-Clarke, 1970 for an early example). Correspondence and audit studies on discrimination can further be divided into two main approaches: The matched or paired, within-subject approach, and the unmatched, between-subject approach. The former design has by far been the modus operandi of researchers on hiring discrimination, but a small number of experiments are designed following the latter approach. Most often, researchers leave the decision on whether or not to send multiple applications per job vacancy unsubstantiated: A choice is simply made, most commonly to send matched pairs. In the few articles where the choice of design is explicitly discussed, some arguments have been formulated – commonly regarding practical and logistical issues or statistical efficiency (see Lahey & Beasley, 2018 for a discussion).

In this paper, I aim to contribute to this somewhat implicit discussion in two ways. First, I aim to systematize and summarize the arguments that have been raised so far in relation to the choice between matched and unmatched experimental designs for detecting discrimination. My point of departure, and example throughout, will be studies of ethnic hiring discrimination, but the arguments in principle extend to other arenas that operate in a similar manner, e.g. housing rental markets. Second, I will more thoroughly address a rarely discussed[1] concern regarding the matched design, that I term the problem of *induced competition*: Forcing a comparison of the treated candidates with a reference candidate potentially influences estimations of discrimination, and this influence is sensitive to the total number of

E-mail address: e.n.larsen@sosgeo.uio.no.

[1] However, the argument is similar to a point raised by Pager (2007) on ethnicity as a way for employers to "break the tie" when candidates are equal in other regards, and to Heckmanös (1998) critique that matched studies may be seen to provoke discrimination where, due to employee sorting, there might not be any in the real world.

https://doi.org/10.1016/j.rssm.2020.100475
Received 27 February 2019; Received in revised form 6 December 2019; Accepted 7 January 2020
Available online 09 January 2020

candidates that apply for a position. In other words, when the experimental applications are just two of a large number of other applicants, the added competition from the reference candidate should be negligible. However, in tighter labor markets where demand is high and fewer candidates apply for each vacancy, the added competition could potentially be more substantial. The unmatched design should in principle be insensitive to these variations. While I will discuss conceptual differences between the approaches as well, this particular concern is unrelated to discussions about which measurement best represents the situations researchers intend to measure: My worry is that paired designs have a built-in sensitivity to the size of the applicant pool size – a number commonly unknown to the researcher, and thus difficult to adjust for. In order to investigate this concern, I develop a computational model that simulates a series of field experiments, varying the type of design, the average number of applicants per vacancy, and other assumptions on employer behavior. My overall conclusion from these theoretical simulations is that the matched design appears to be sensitive to the average applicant pool size in a way that the unmatched design is not, which can be concerning when comparisons between labor markets, points in time or different occupations are made.

My main research questions can thus be summarized as follows:

- What differences, both substantially and practically, should researchers take into account when deciding between matched and unmatched correspondence study designs?
- Do matched and unmatched designs differ in their sensitivity to the average size of the applicant pool in correspondence studies, and what implications does this sensitivity have for the design and interpretation of correspondence studies?

2. Reasons and reservations in choice of design

As noted in the introduction, audit- and correspondence testing has received growing attention among social scientists interested in studying discrimination and labor market inequality along various dimensions. The overwhelming majority of studies on discrimination follow the matched, within-subject design, where pairs of equally qualified testers or applicants differing only in the treatment itself are submitted to the same job vacancy (for historical and methodological overviews, see e.g. Bovenkerk, 1992; Cherry & Bendick, 2018; Neumark, 2018; Shadish, Cook, & Campbell, 2002). Taking studies of ethnic hiring discrimination as a case, Table 1 shows an overview of most correspondence and audit studies on ethnic hiring discrimination conducted in the past decade.

Out of these 35 publications, only six are based on studies that follow an unmatched design - however, these examples are fairly recent and could be indicative of a growing trend. A possible explanation for the historical popularity of the matched design is the ties to activist research traditions and research for enforcement purposes, where producing convincing evidence of discriminatory practice among specific employers is seen as crucial (see Cherry & Bendick, 2018 for a discussion). As will be discussed in the following, the matched design has several other attractive features that can explain its historical appeal, but the option of not matching has perhaps received an unwarranted lack of attention. Before proceeding with what is the main contribution of this article, the issue of induced competition in matched designs, I will review a range of factors relevant to the comparison of the two design approaches that have been discussed in previous studies on the topic.

2.1. Statistical efficiency

The statistical properties of matched and unmatched designs have most thoroughly been approached by Vuolo, Uggen, and Lageson, 2016; Vuolo, Uggen, & Lageson, 2018). Statistical efficiency – that is, needing smaller samples to reach satisfactory levels of statistical power – is crucial for field experiments on hiring discrimination, as they are often costly endeavors. Commonly, it is believed that matched designs are more statistically efficient than unmatched designs in general. Taking this belief as their starting point, the authors convincingly argue that whether or not this belief holds true depends on the level of concordance, i.e. the proportion of cases where both applications in a pair receive the same response. When concordance is above 0.5, the matched design remains more statistically efficient in most cases. Conversely, when concordance is lower, the unmatched design has an efficiency advantage: "for matching to matter, the experimental unit (e.g. employers) must respond at least somewhat similarly to the matched pairs" (Vuolo et al., 2018, p. 131). However, it should be noted that even if this is the case, the degree of concordance is typically high in studies of hiring discrimination, as overall callback rates tend to be low. In practice, this means that in most cases, the matched approach might still be the most efficient choice when testing for hiring discrimination. While the level of concordance is generally unknown for a given correspondence test, it can be estimated by drawing on previous experiments in the same context, or better yet, through pilot studies.

Related to the question of statistical efficiency is that of access to experimental *units*, e.g. suitable vacancies for a hiring discrimination

Table 1
Audit- and correspondence studies on ethnic hiring discrimination, 2010-2019.

Author (Year)	Matching (number per vacancy)
Jackson (2009)	Unmatched paired (2)*
Duguet, Leandri, L'Horty, and Petit (2010)	Yes (2)
McGinnity and Lunn (2011)	Yes (2)
Oreopoulos (2011)	Yes (4)
Andriessen, Nievers, Dagevos, and Faulk (2012)	Yes (2)
Agerström et al. (2012)	No
Berson (2012)	Yes (2)
Booth, Leigh, and Varganova (2012)	Yes (4)
Capéau, Eeman, Groenez, and Lamberts (2012)	Yes (2)
Carlsson and Rooth (2012)	Yes (2)
Derous, Ryan, and Nguyen (2012)	Yes (4)
Drydakis (2012)	Yes (2)
Jacquemet and Yannelis (2012)	Yes (3)
Kaas and Manger (2012)	Yes (2)
Maurer-Fazio (2012)	Yes (2)
Edo, Jacquemet, and Yannelis (2013)	Yes (6)
Pierné (2013)	Yes (6)
Arceo-Gomez and Campos-Vazquez (2014)	Yes (8)
Blommaert, Coenders, and van Tubergen (2014)	No
Bursell (2014)	Yes (2)
Galarza and Yamada (2014)	Yes (4)
Baert et al. (2015)	Yes (2)
Decker, Ortiz, Cassia, and Hedberg (2015)	Yes (2)
Nunley, Pugh, Romero, and Seals (2015)	Yes (4)
Gaddis (2015)	Yes (2)
Agan and Starr (2016)	Yes (4)
Baert and Vujić (2016)	Yes (2)
Darolia, Koedel, Martorell, Wilson, and Perez-Arce (2016)	Yes (2)
Kang, DeCelles, Tilcsik, and Jun (2016)	No
Lee and Khalid (2016)	Yes (4)
Midtbøen (2016)	Yes (2)
Weichselbaumer (2016)	Yes (2) and No
Baert, Albanese, du Gardein, Ovaere, and Stappers (2017)	Yes (2)
Birkelund, Heggebø, and Rogstad (2017)	Yes (2)
Vuolo, Uggen, and Lageson (2017)	No
Koopmans, Veit, and Yemane (2018)	No
(Lancee et al., 2019a)	No

Note: Only one publication cited for each individual correspondence study for clarity. Overview obtained from Baert (2018); Zschirnt and Ruedin (2016) and Vuolo et al. (2018) with more recent and other studies added.

* Jackson's study uses pairs but forgoes matching, and is discussed further in the concluding section.

experiment. In this case, the matched approach offers a clear advantage. In order to reach the same number of observations, a paired designs requires half the number of suitable job vacancies. This is especially critical when conducting field experiments in smaller labor markets, where there are fewer available vacancies at any given time[2] . In cases where the degree of concordance is high – i.e. in many studies of hiring discrimination – this advantage of the matched approach becomes even clearer.

2.2. Multiple treatments

While statistical efficiency is often implicitly perceived as a strength of the matched approach, as scholarly interest has moved towards exploring mediating and interacting factors in hiring discrimination, the ability to randomize across multiple experimental treatments is raised as an important strength of the unmatched approach. By employing the same logic of a full factorial survey experiment, an unmatched field experiment allows researchers to randomize – in principle, barring concerns of statistical power - as many treatments as she likes. Of course, careful consideration must be taken when conducting power analysis for these experiments, in particular if interactions between different treatments are of interest. An illustrative example would be if a study sets out to not only measure ethnic discrimination of a particular minority group, but of several: In this case, the full factorial unmatched design allows for easily implementing multiple treatment conditions without having to send an excessive number of fictitious applicants to the same employer. However, as mentioned above, this in turn requires larger number of experimental units, i.e. vacancies. It is also possible to exploit both within- and between-pair variation in matched designs, allowing for the inclusion of multiple treatments without increasing the number of applicants sent to a single employer, by simultaneously varying two treatments (i.e. race and educational credentials) within each pair (see e.g. Gaddis, 2015 for an example and discussion. However, this approach removes the desired strict comparison of candidates equal on all but one dimension of interest: whether or not this is problematic depends on ones view of what precisely constitutes measuring discrimination. This will be discussed further in Section 2.4

2.3. Risk of detection and ethical considerations

Following the previous point, there are potential advantages to not relying on sending multiple fictitious applications to the same job vacancy. The first reason revolves around the risk of detection, which of course is something researchers strive to avoid. Even submitting two applications that are meant to only differ in one aspect, e.g. the name of the applicant poses a design challenge regarding the risk of detection: This is conventionally solved by creating two (or more) templates which differ only cosmetically. Further, these templates are randomly assigned orthogonally to the treatment, which allows researchers to control for any effects of the templates in analysis. However, if one wants to vary multiple treatments simultaneously, the design (and analysis) becomes exponentially more convoluted and challenging to implement – and the risk of detection grows. Detection is also only observable by the researchers in certain overt cases, so any potential bias stemming from high detection rates could go unnoticed. An illustrative example of this problem can be found in Weichselbaumer (2015, 2016), where the experimental design was in fact changed from matched to unmatched due to severe issues with detection.

The second aspect of being unconstrained by the need to create identical-yet-different applications is related to ethical concerns. Field experiments are intrusive, and rely on subverting the principle of informed consent in order to preserve external validity (for thorough discussions, see Riach & Rich, 2004b and Zschirnt, 2019). Usually, researchers preparing a correspondence study face ethics boards – either institutional or governmental – and rely on a series of arguments to justify this subversion. These will typically appeal to the social importance of the topic weighted against the intrusion involved in making employers waste time on evaluating fictitious applicants. The magnitude of this intrusion is directly related to the number of applicants one sends to the same employer, so it follows that single-application factorial designs are more readily justifiable from the perspective of research ethics (see Lahey & Beasley, 2018, p. 91 for a similar argument).

2.4. Substantial considerations in measuring discrimination

From a conceptual and theoretical point of view, one could pose the following question: *Do matched and unmatched approaches measure different things?* In other words, how do we interpret the results from correspondence studies, and does interpretation depend on the design? The answer to this question relies on how one conceptualizes discrimination. Blank, Dabady, and Citro (2004) suggest a two-component social science definition of discrimination: "(1) *differential treatment on the basis of race* that disadvantages a racial group and (2) *treatment on the basis of inadequately justified factors other than race* that disadvantages a racial group (differential effect)." (p. 39). Correspondence studies are designed to measure the former by isolating the causal effect of race.

Commonly, analysis of matched and unmatched designs are conducted in a similar manner: by comparing aggregate callbacks of treated and reference applicants, often expressed as callback ratios. With an unmatched design, this number expresses how much less (or more) likely a treated candidate is to receive a callback, relative to a reference candidate. For the matched design, however, such an interpretation needs to be qualified by adding that we measure a slightly different probability for the treated candidates: That is, the probability of receiving a callback in face of – among others - an equally qualified competing candidate *without* the treatment. Which number to be preferred is a substantial and conceptual rather than a practical or methodological question, and in the following, I will briefly review the discussion so far, before arguing against a clear substantial distinction between the measurements.

Central to the question of how the designs differ substantially is how one defines discrimination. One position found in the literature is the view that matching is a strict requirement for measuring discrimination. According to this view, unmatched designs measure something conceptually different, as expressed by e.g. Riach and Rich (2004a), p. 471), who use the term 'preferential treatment'. The authors cite Heckman and Siegelman (1993)[3], and conclude that that discrimination can only occur when an individual employer is confronted with a need to choose. They further argue that "…no illegal activity is detected by this [unmatched] procedure." (p. 471). In my view, there are two objections to this strict stance on what constitutes discrimination. First, a core element in definitions of (direct) discrimination as a phenomenon

[2] When collecting data for our own project (Lancee et al., 2019a, 2019b), this weakness of an unmatched design became clear: Achieving the same number of occupations in Norway and the Netherlands as in Germany or Spain turned out to be more challenging than anticipated, and led us to implement additional occupations in the former countries. This, in turn, needs to be taken into account if one wishes to compare national contexts.

[3] The Heckman and Siegelman (1993) quote in question is the following: "Discrimination exists whenever two testers in a matched pair are treated differently in the aggregate or on average" (p. 198), which is taken as support for the view that unmatched designs do not measure discrimination. I do not agree that Heckman & Siegelman express a view that discrimination *only* exist when the above is the case. The statement is made in regards to how one interprets asymmetric treatment in *matched* audit studies. Thus, I do not find this a convincing argument in itself for why unmatched designs are substantially unable to measure discrimination.

studied by social scientists is precisely differential treatment based on some irrelevant trait or category; this is also in line with the definition proposed by Blank et al. (2004) discussed above. They simultaneously argue for a distinction between legal and social science understandings of discrimination, where the latter should be broad enough to "include a range of behaviors and processes that are either not explicitly unlawful or not effectively prohibited because of difficulties in measurement or proof" (p.41). Thus, one can argue that a clear distinction between discrimination and differential treatment has limited relevance when enforcement is not an explicit focus. Second, one can argue that even when conceding that only specific acts where a minority applicant is treated worse compared to an equally (or less) qualified majority applicant constitute discrimination, unmatched designs still measure these actions: If there are aggregate systematic differences in callback rates between treated and untreated candidates, even discrimination in the strict sense must have occurred *in some hiring processes* in order to produce these differences. The problem of being unable to identify these occurrences, however, remains.

Cherry and Bendick (2018), in a similar vein, correctly notes that unmatched designs prevent researchers from identifying individual decision-makers or firms, and further concludes that "…unpaired audits describe a *villainy without villains"* (p.55), but that unpaired auditing studies "…would be acceptable for scholarly publication." (Cherry and Bendick, 2018). They do not, however express a strong stance on whether or not unmatched studies measure discrimination (if not with identifiable villains), and their main concern is how the findings from such studies can be mobilized for activist and enforcement purposes, i.e. whether or not they produce convincing evidence of discrimination. This issue is, in my view, separate from the substantial debate on what the designs measure. It is likely true that, regardless of how one conceptualizes discrimination, the inability to identify specific discriminatory actors makes unpaired correspondence studies *de facto* unsuitable for enforcement purposes. Whether or not this is problematic thus depends on the motivation of the study. In correspondence tests made for research purposes (as opposed to legal enforcement), the primary interest commonly lies in detecting discrimination as a market-level phenomenon, not in identifying firms that acted discriminatory per se. Thus, unpaired designs should produce satisfying measurements. Finally, it is possible to contest the claim that *matched* designs are able to identify individual discriminatory actors. Midtbøen (2013) notes that "…the single act of choosing one candidate in favor of another cannot be defined as direct discrimination because it could be the result of a coincidental preference for one out of two equally qualified job applicants" (p.29). This issue of separating randomness from direct discrimination has also been noted by Pager and Western (2012), p. 233) and Heckman and Siegelman (1993), p. 198).

As noted, correspondence studies are primarily concerned with describing discrimination as a macro phenomenon, and in these cases, the inability to (in theory, but keep in mind the argument above) identify specific firms is not a concern. However, there are notable exceptions to this: Pagerös (2016) imaginative study on the potential long-term consequences of discrimination for firms themselves would not be possible based on data from an unmatched design, as her analysis relies on identifying which particular firms expressed tendencies towards discrimination. Bonoli and Fossati (2018) exploit information about the cases where the *treated* candidate is preferred, discovering interesting patterns of minority preference among high skilled professional occupations. A similar point in favor of a matched design is the ability to exploit additional information about other aspects of the hiring process such as the *order* in which applicants received callbacks, more subtle differences in how applicants are treated (e.g. differences in politeness or brevity and the length of job interviews for audit studies, see for instance Bendick, 1996; Bendick, Jackson, Reinoso, & Hodges, 1991; Bendick, Rodriguez, & Jayaraman, 2010; Ghumman & Ryan, 2013 and Pager, Bonikowski, & Western, 2009). Such information is unavailable in unmatched designs, but is possible to proxy by comparing group averages when statistical power permits it[4] .

Further, one can discuss validity of the actual situation produced in the two design approaches. The logic of the matched design is to control all elements of the (early stages of the) hiring situation by presenting the employer with two (or more) identical candidates differing solely on one dimension. However, as Agerström, Björklund, Carlsson, and Rooth (2012)) correctly points out, such a situation might be rare and unrealistic – even if it does reveal discriminatory behavior (see also Heckman, 1998). They defend their choice of an unmatched design by stressing that:

> …although most experiments aim to reduce random noise, the present study instead aimed to induce a background of highly realistic noise. To the extent that we then find statistically reliable effects of the variables we are interested in (warmth, competence and ethnicity), we can be confident that they are robust over a wide range of variables that are likely to vary in a real-life context, thus greatly increasing the generalizability of our findings (Agerström et al., 2012, p. 361–362).

Even without varying multiple characteristics to create a "background of highly realistic noise", the single-application simply pits the treated candidates against real-world competitors that would be faced in a real job seeking process – not against the added competition of another fictitious identical candidate (see also Carlsson, Fumarco, & Rooth, 2015 for a similar argument). Finally, there is also some evidence of spillover effects in correspondence studies, an issue raised by Phillips (2016): That is, the notion that a fictitious application's callback probability is influenced by employer perceptions of the applicant pool as a whole.

3. The induced competition of matched designs

The above points lead me to what is the central concern of this paper, namely that in matched designs, the probability for the treated candidate to receive a callback can be directly influenced by the added competition of the reference candidate. Whether or not this is the case depends on a few assumptions regarding how the invitation process actually unfolds, and requires us to think carefully about how we imagine it to function. This should only be the case if employers have some kind of limit to how many applicants they are willing to invite to an interview: by adding a more desirable reference competitor (assuming that employers discriminate), one decreases the probability for the treated candidate to receive one of the available interview invitations. One might argue that this in and of itself is not a problem – after all, even a case where the minority applicant otherwise would have been invited to an interview if it were not for the reference competitor trumping them is still clear evidence of discrimination. So far, this mechanism is similar to the issue of matched designs making ethnicity a mere tie-breaker (see Heckman, 1998 and Pager, 2007, p. 116). My main concern, however, is that this effect of induced competition is sensitive to the total number of applicants for a given vacancy. In cases where an employer receives few applications, the paired design would potentially produce higher (but not necessarily biased or incorrect) estimates of discrimination if the reference competitor nudges the treated candidate out of the shortlist for an interview. However, if an average vacancy receives a large number of applicants, the effect of the added reference competitor would be negligible. Thus, while the question of whether or not matched or unmatched designs produce essentially truer estimates of discrimination remains a topic for discussion, I argue that the matched design is potentially sensitive to the

[4] As with differences in callbacks, one cannot identify which specific firms produce the differences, but any systematic difference between groups (in either e.g. contact time, politeness or interview lengths), assuming proper randomization, must be produced by the treatment itself.

average number of applicants a vacancy receives – in a way that the unmatched design is not. As the total number of applicants an employer receives is usually unknown to the researcher, this sensitivity threatens any comparison one wishes to make; be it across occupations, labor market contexts or points in time.

How would one test for this sensitivity empirically, and determine its potential magnitude? The true empirical test requires three criteria to be satisfied: First, there must be an underlying treatment effect – i.e. discrimination – for the study to measure. Second, there needs to be variation in experimental design, i.e. between matched and unmatched applications. This would be difficult to motivate as its own study, but perhaps a possible addition to an already planned correspondence test. Finally, the number of applications a given vacancy received must be known to the researcher, or at least proxied in a satisfying manner. Whether this is possible depends on whether such information is offered by the job search platform used, or it could possibly be surveyed directly in retrospect – which itself would be a costly venture. If all of these criteria are met, one could predict measured discrimination by design (matched or unmatched) interacted with the applicant pool size. If this interaction term is significant, showing that matched designs estimate higher discrimination when applicant pools are small but that this is not the case for unmatched designs, one would have evidence of the sensitivity I describe. Thus, an empirical meta-analysis attempting to identify this sensitivity would require both variation in experimental design (of which there is little, as the overwhelming majority of correspondence studies on hiring discrimination are matched – see Table 1) *and* knowledge of applicant pool sizes (which is rarely possible to obtain). One fundamental issue remains that obstructs any clear empirical test for the induced competition sensitivity: that labor market tightness, i.e. applicant pool sizes, is *itself* thought to influence employers' discriminatory behavior. Even though the empirical evidence is inconclusive (see Baert, Cockx, Gheyle, & Vandamme, 2015; Carlsson, Fumarco, & Rooth, 2018), there are theoretical reasons to expect employers to be less discriminatory when need of labour is high. This means that a meta-analysis identifying trends of higher measured discrimination in matched studies when applicant pools are small, will have trouble disentangling whether such a pattern is due to the mechanism of induced competition that is the topic of this paper, or due to the substantial effect of labour market tightness on employer behaviour.

The above discussed reasons prevent an empirical investigation of the potential effect of induced competitions in matched correspondence tests, barring conducting a large-scale correspondence study randomizing both design (matching or not) and obtaining knowledge on applicant pool sizes. A different approach to estimate its potential severity is to approach it through simulation. The remainder of this paper is dedicated to computationally mapping out the potential effect by creating a simulated meta-analysis of correspondence tests, varying applicant pool sizes, design choice, and exploring their interplay in affecting discrimination estimates.

3.1. The effect of competition: a simulation exercise

The aim of this exercise is to quantify the potential sensitivity incurred by the matched design through producing a large number of simulated studies where all parameters of interest can be manipulated. In the following, I will outline the simulation procedure.

Initially, a dataset of 2000 of observations is generated, and a treatment is randomly allocated to half of the sample. These observations represent fictitious experimental applicants in a field experiment. If the simulated study is matched, pairs of applicants are submitted to the same vacancy, and treatment randomization occurs within pairs. In the unmatched, single-application variant, each applicant is submitted to an individual vacancy. The size of the applicant pools for vacancies in the given simulated experiment is then specified: This number is varied from a minimum of 5 to a maximum of 50: When applicant pool sizes average below 5, basic assumptions (which will be discussed later) will drastically affect outcomes[5] (even if such cases most definitely exist in real life). The upper limit of 50 is set based on converging results, which will be shown further in the analysis.

For each vacancy, a recruitment process is simulated. This is an idealized setting where a pool of applicants is constructed (with a size defined as described above), consisting of our applicant(s) of interest from the generated correspondence study, and a list of other applicants representing real-world applicants outside of the experiment. Each of the applicants, including the experimental ones, are assigned a random value from a uniform distribution ranging from 0 to 1 representing their evaluation score from the employer (thus making no assumptions regarding where in the distribution of other applicants our fictitious ones lie). Treated applicants (one of the two in the matched approach, and either one or none in the unmatched approach) are subtracted a predefined treatment effect[6] . This implementation allows for heterogeneity in employers' preferences; sometimes, the treated candidate will be evaluated higher than the control candidate will due to random variation in their initial ratings.

Finally, all applicants applying to a specific vacancy are ranked according to this random rating. The highest-scoring proportion of the applicant pool size defined by the variable 'baseline proportion called back' are invited to an interview[7] . This number, in practice, represents the overall probability for a control-group candidate to receive a callback, i.e. the baseline callback rate. The following example illustrates the complete procedure that is simulated for each vacancy. Let us assume that the experiment is matched, so that both a reference and a treated candidate are submitted to the vacancy in question. The applicant pool size is set to 8, which means that 8 other applicants will be applying. The reference candidate receives a random rating of 0.7, while the treated candidate receives a random rating of 0.8, which is then subtracted -0.1 resulting in a rating of 0.7. Among the 10 candidates, this should, due to the random uniform rating of the other candidate, make it likely that both candidates are among the top half of the applicant pool (as the mean of the other, external candidates will be 0.5), potentially resulting in both candidates receiving a callback (even though the reference competitor was rated slightly higher)[8] . This procedure is repeated for every vacancy in the simulated study. Table 2 shows an overview of all variables manipulated in the simulation exercise, and occasionally refers to three scenarios (A, B and C) which will be discussed in the next section.

The end result is a dataset of a simulated field experiment where the underlying treatment effect and the average applicant pool size is known. Randomized datasets are then simulated over a large number of repetitions, varying the design (matched vs unmatched) and the applicant pool size. As researchers are commonly interested in *relative* differences, e.g. expressed as odds ratios, the callback ratio between treatment and control groups is stored as the output from each repetition[9] . The R code for the simulations are available in the

[5] E.g. whether one assumes that employers always invite at least one candidate or not, in extreme cases where only the fictitious applicants apply.

[6] Note that this penalty to the abstract random rating is simply a way to operationalize an underlying treatment effect, and does not necessarily correspond to any specific measurement. How much of an impact this subtraction has depends on the other parameters, but is not of central interest here.

[7] For scenario A, every applicant that falls above a threshold of 0.5 is invited to an interview instead of a highest proportion. This means that many candidates are invited when applicant pools are large, which is likely unrealistic, but keeping the number relative to the pool size keeps the underlying discrimination coefficient constant across applicant pool sizes, which in turn makes comparison between designs clearer.

[8] Similarly, in scenario A, both candidates would in this case be invited, as both score above the 0.5 rating threshold.

[9] I have also repeated the analysis with a coefficient estimate from a linear probability model as the output per simulated study, and the patterns remain

Table 2
Overview of variables in simulation procedure.

Variable	Value range in simulation	Explanation
Number of observations	Held constant at 2000 (50 % with treatment).	The total number of applicants submitted. In unmatched design, this is equal to the number of hiring processes. In the matched design, it is equal to twice the number of hiring processes/vacancies. The value of 2000 was chosen to ensure sufficient statistical power for each simulated study.
Applicant pool size	5 – 50.	The number of other, non-experimental applicants in the simulated correspondence study. These are randomly rated by employers between 0 and 1.
Treatment effect	Held constant at -0.1. This means that reference candidates on average are rated 0.5, while treated candidates on average are rated 0.4.	A specified value that is subtracted from treated applicants' rating score (which in turn is a random number between 0 and 1).
Baseline proportion called back	Scenario A: Not used. Scenario B: 50 %Scenario C: 50 %	The proportion of all applicants for a given vacancy that receive a callback.
Lower-limit rating threshold	Scenario A: 0.5. Scenario B: Not used. Scenario C: 0.5	A value representing the lower limit rating for an applicant to receive a callback, regardless of where they fall in the ranking order of the applicant pool. This allows for employers to invite no candidates, if they all fall below the threshold.

supplementary material (S1). In the following, I visualize the results of a meta-regression of simulated field experiments[10] under three different scenarios with different operationalizations of employer behavior.

4. Analysis of simulated experimental data

My overarching assumption is thus the following: When the average job vacancy receives a small number of applications, the paired application procedure produces different discrimination estimates because the researcher induces additional competition. More specifically, I expect the paired approach to produce higher estimates of the treatment effect when the application pool size is small. I expect this effect to diminish as the mean application pool size grows, and in addition, that the presence of the effect relies on how we assume the recruitment process to function. To shed light on the latter point, I have constructed three experimental scenarios, each representing sets of assumptions regarding the recruitment process. In the first scenario (A), employers simply invite candidates that exceed a certain rating threshold – regardless of the relative composition of the given applicant pool. In this case, I expect the designs to produce similar results, and the estimates should be unaffected by the applicant pool sizes. In the second scenario (B), employers rank the submitted applicants and respond positively to a specified top proportion. In this setting, I expect the experimental design to matter, as the induced competition by adding the reference candidate directly influences the probability for the treated candidate to receive a callback – unless the applicant pool is sufficiently large. Finally, the third scenario (C) represents a combination of the two preceding scenarios: Employers initially shortlist the top specified proportion of applicants, but also have a limit to how low an invited applicant can be rated. In scenarios A and C, I thus allow for the possibility that the employer does not invite any candidates. In every scenario, the number of observations for each study is 2000 and the actual treatment effect, modeled as the decrease in rating received by the treated candidates, is -0.1. Thus, each simulated study is sufficiently powered to detect the effect, and statistical significance is disregarded in the analysis.

Empirical analysis of actual hiring processes is required to determine which of these scenarios are more realistic, and we can expect them to vary widely between national contexts, industries, sectors and firms. However, I argue that the three idealized scenarios above should capture a range of plausible processes sufficient to address the concern that is the topic of this paper. I encourage the reader to determine which scenario is more relevant to their specific setting of interest.

4.1. Scenario A: No relative competition, absolute ratings threshold

In the first scenario, we specify a lower rating threshold under which employers are not willing to hire applicants. Substantially, we assume that employers are simply looking for candidates that qualify: The ranking order of the candidates in the pool does not matter, and only candidates randomly rated above the threshold are invited irrespectively of the applicant pool for the given vacancy. This implies that the treated candidates are less likely to be invited compared to the control candidates, and this should be the case regardless of experimental design. In practice, this scenario in part serves as a robustness check for whether or not the simulation yields expected results, as well as a baseline for later comparison.

Fig. 1 shows the result of 10 000 simulated experiments under the assumptions discussed above. The vertical axis shows the callback ratio, i.e. the callback rate of the control group divided by that of the treated group, produced by each simulated correspondence study (each point represents a single simulated correspondence study). The x-axis shows applicant pool size in each given experiment, and colors indicate study design. A loess regression slope with 95 % confidence intervals is added for clarity. As expected, when employers operate on absolute rating thresholds, both designs produce identical estimates of the underlying treatment effect (which equals a callback ratio of approximately 1.25), and these estimates are unaffected by the average applicant pool size.

4.2. Scenario B: relative competition, no ratings threshold

In the following scenario, I specify the proportion of applicants for a given vacancy who receive a response: Substantially, this means that we assume that employers only care about the relative composition of the applicant pool for the given position, not their rating in absolute terms. The following example provides a simple illustration. Assume we specify a paired design, a baseline response proportion of 0.2, and a vacancy with a given applicant pool of size eight, which means that eight additional applicants will apply. The ten total applicants (including our treated and untreated candidate) are then randomly rated between 0 and 1, the treated candidates receives a penalty corresponding to the assumed treatment effect, and finally the top five candidates receive an invitation: These final five might or might not contain our experimental applicants, but are more likely to contain the untreated applicant. In this case, the experimental design should matter when applicant pools are small, as the paired design adds a reference group competitor,

(*footnote continued*)
unchanged (however, less intuitively interpretable).

[10] While the standard procedures for paired and unpaired design would differ somewhat (i.e. cluster-corrected standard errors by vacancy in paired designs), this is irrelevant for the present study: All estimates are significant at the 1% level due to the high N of each individual simulated experiment, and the estimates themselves are the main interest – not their associated statistical power.

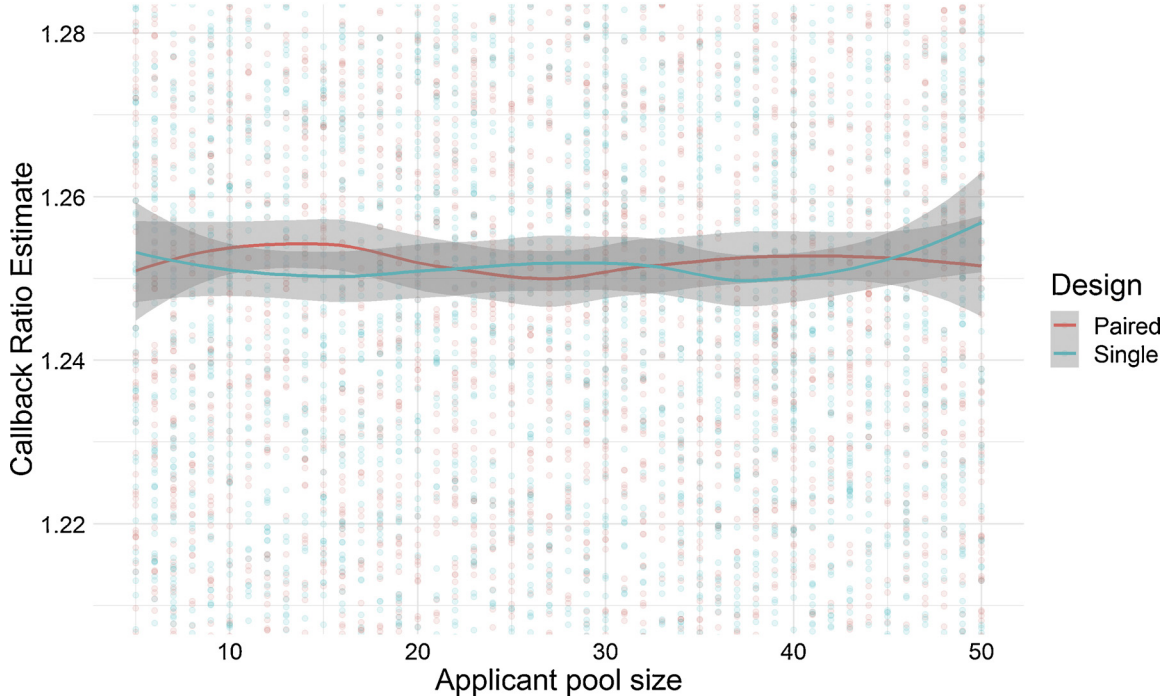

Fig. 1. Scenario A. Callback ratio estimates from 10,000 simulated correspondence studies by mean applicant pool size. Colored lines indicate fitted loess regression slope with 95 % confidence intervals.

lowering – on average – the probability that the treated candidate receives an invitation. I specify the baseline proportion invited to be 50 %, e.g. that the highest-rated half of all applicants are invited for a given vacancy[11] .

Fig. 2 shows the hypothesized effect of induced competition in the paired experiments: The paired design produces larger callback ratio estimates when applicant pools are small, but the results converge when pool sizes approach 30. Again, whether the estimates align with the predetermined "true" treatment effect (shown to be approximately 1.25 in scenario A) is not of primary interest – but whether the designs are sensitive to changes in the applicant pool size. Under these assumptions, the paired design exhibits the expected sensitivity, while the single application design produces consistent estimates across applicant pool sizes. Note that the paired design produces slightly higher estimates across all applicant pool sizes as well, but that this difference is negligibly small (and not a concern in those cases it is consistent across pool sizes).

4.3. Scenario C: relative competition, absolute ratings threshold

The final scenario is a combination of the above-discussed two, as this is arguably more realistic. While scenario B forces employers to invite the top half of applicants regardless of absolute rating scores, I modify this assumption in the following. Employers still aim to invite a certain top proportion of the candidates who applied, but they additionally have a lower-bound threshold under which they are not willing to invite candidates (specified to 0.5, thus eliminating the bottom-half of all applicants). This allows us to part with the unrealistic assumption of scenario 1, by letting employers abstain from inviting any candidate if there are few and none are qualified (Fig. 3).

[11] Note that this assumes that employers adapt the number of invitations to the total size of the applicant pool. It is indeed possible that employers have an upper limit for how many invitations they are willing to provide, but modelling the number of invitations as a proportion has the advantage of keeping the underlying estimate similar across pool sizes. This, in turn, makes the results easier to compare and interpret.

The pattern is similar to that of the second scenario, but interestingly, *both* designs exhibit some sensitivity to the applicant pool size: Before the applicant pool size increases beyond 20, the two designs produce higher callback ratio estimates – however, the effect is more severe for the paired design. The sensitivity of the paired design can be explained by the induced competition of the reference candidate, but the case of the unpaired design is more puzzling. A possible explanation is that this implementation of the invitation procedure simply produces larger callback gaps at small applicant pool sizes regardless of the design, and that the issue of induced competition additionally affects the paired designs.

Thus, while both designs would be expected to produce slightly varying estimates depending on the applicant pool, the paired design is more sensitive. How substantial this sensitivity is depends on how actual hiring processes function in detail, as illustrated by scenarios A, B and C. In the case of scenario B, the distortion changes estimates from approximately ratios of 1.25–1.285 under the paired design, which might not in itself be a large difference. However, if we imagine a study comparing two different industries where labor market tightness differs, the paired design runs a larger risk of either exaggerating or downplaying any actual differences in discrimination - purely as an artefact of the design.

Finally, it is worth noting that in all three scenarios, I have chosen to model employer behavior and preferences as unaffected by the applicant pool size itself. In scenario A, all qualified (i.e. those falling above the threshold) are invited regardless of numbers. In scenario B, the top 50 % are invited, again, regardless of numbers, while scenario C is a combination of both. This constancy of employer preferences regardless of applicant pool sizes is a strong assumption, as there is reason to believe that employers adapt their behavior with regards to the number of applicants as discussed in Section 3. I have chosen to retain this simplification, as it allows me to isolate the potential effect of induced competition and preserve clarity and interpretability of the results, at the cost of realism. It would be possible to model a wide range of different employer behaviors for different applicant pool sizes, but in principle, wherever there is *some* form of ranking involved and competition for a finite number of callbacks in employer decisions, the

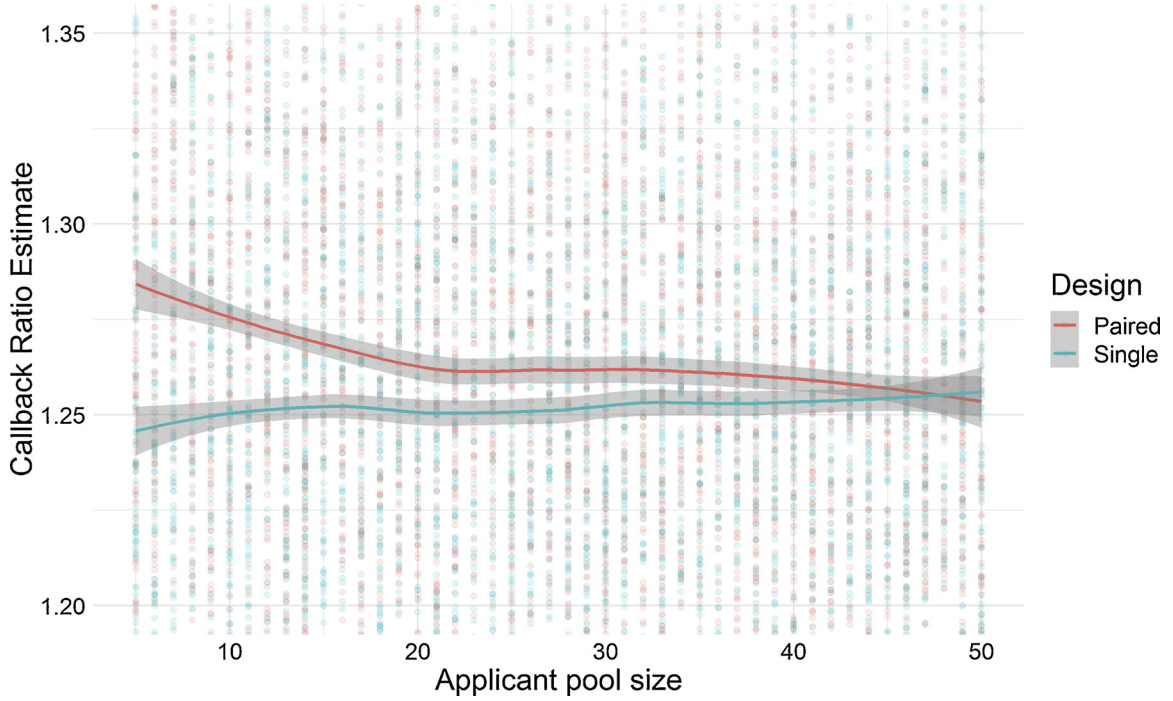

Fig. 2. Scenario B. Callback ratio estimates from 10,000 simulated correspondence studies by mean applicant pool size. Colored lines indicate fitted loess regression slope with 95 % confidence intervals.

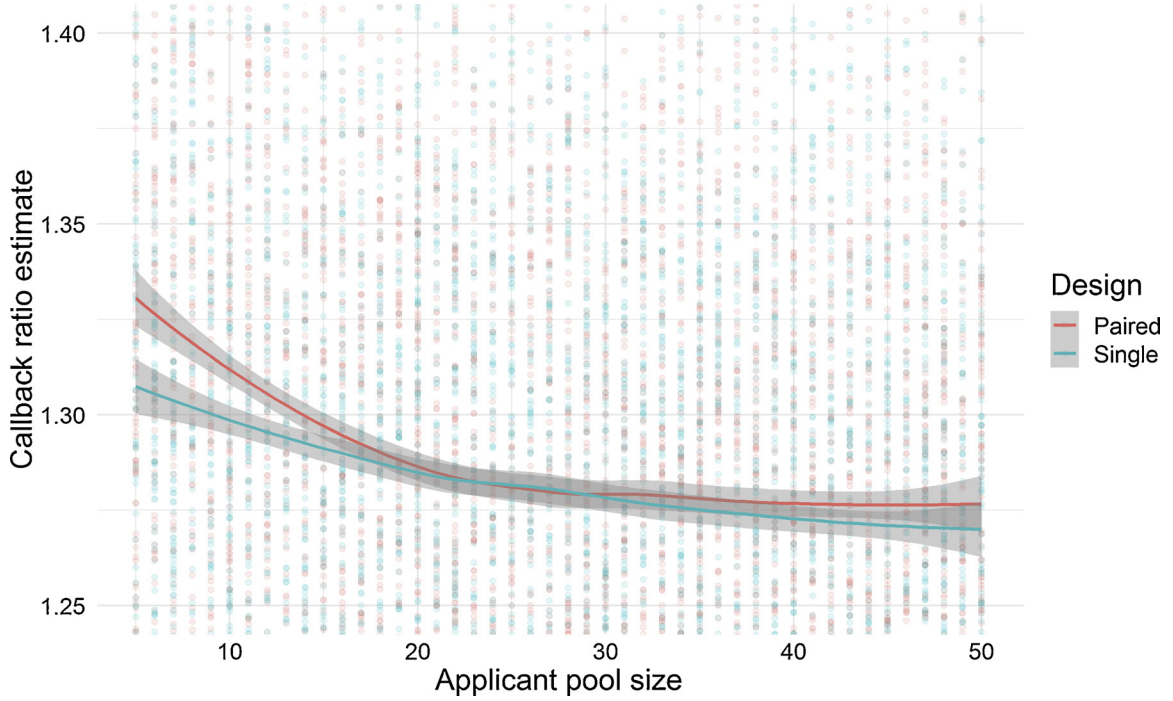

Fig. 3. Scenario C. Callback ratio estimates from 10,000 simulated correspondence studies by mean applicant pool size. Colored lines indicate fitted loess regression slope with 95 % confidence intervals.

effect of induced competition might constitute an issue – regardless of whether employers discriminate more or less at different applicant pool sizes.

5. Discussion and summary

I have shown the potential distortion caused by matched pair designs in correspondence studies through the additional induced competition they involve. What are the practical implications of this finding? The impact of this distortion for a given study or a given comparison between studies depends on multiple parameters: The callback rate of the reference category, the magnitude of the difference in callbacks, and the practices of employers in ranking and inviting candidates for interviews. Thus, how problematic this would be depends on the particularities of a given study, but can in principle have serious implications for studies aiming to provide comparative

evidence. If labor scarcity varies while underlying discrimination rates are equal, a significant difference could be detected in a harmonized paired design – but in reality purely be an artifact of the study design. Comparing two unpaired experiments, however, does not bear with it this risk to the same extent, as the underlying treatment effect is more similarly estimated across all applicant pool sizes. In summary, the paired design is potentially sensitive to the mean number of applicants per vacancy in a way that the unpaired design is not. There are two main issues with providing an empirical test for this sensitivity. First of all, labor market tightness can potentially impact discrimination rates by itself, making it difficult to separate this substantial effect from the design effect of induced competition. Second, researchers usually have no direct knowledge of the actual applicant pool sizes for correspondence studies, so it is difficult to correct for the sensitivity post-hoc. My suggestion for researchers planning to implement a correspondence test is to try to obtain this information before going into the field. Since the effect of induced competition I describe only appears at relatively small applicant pool sizes, one should be safe to choose either a matched or unmatched approach if there is reason to believe that vacancies typically receive large numbers of applicants. If, however, one suspects that there might be contexts in the study where applicant pools on average are very small (i.e. occupations with extremely high demand for labor), one should consider opting for an unmatched approach to safeguard against the effect of induced competition. Information about applicant pool sizes can also be approximated using supply/demand ratios or administrative data if available.

There is an additional facet to the induced competition I argue affects measurements of discrimination in matched designs, which I have not taken into account in this stylized simulation exercise, and that is the *qualitative* dimension of applicant profiles. In these simulation models, I imagine applicant evaluation as a highly abstract, unidimensional process represented by a single value which employers base their decisions on. However, one can easily imagine a more complex process, where applicants – according to their particular credentials and labor market experience – are ranked along different dimensions. In this case, the induced competition from the paired design not only functions through adding another competing applicant to the pool: Indeed, the added competitor (depending on design of the applicant templates) also shares the exact working experience and qualifications of the treated applicant. It is easily imaginable that some applications achieve success in evoking responses through their uniqueness or their potential for filling a niche. In a pairwise design with a small applicant pool, this niche potential is reduced – potentially further influencing estimates of ethnic discrimination. Again, the magnitude of this effect would depend on the average number of applicants per vacancy. Future correspondence studies where the design itself is varied, i.e. where some applicants are submitted in pairs and others are not, would go a long way in providing empirical evidence of how design choice potentially impacts discrimination estimates when applicant pool sizes vary.

The potential sensitivity to applicant pool sizes is indeed an argument in favor of choosing unmatched designs in cases where one suspects applicant pool sizes to vary greatly between contexts and occupations, but there are multiple factors to take into account when deciding on a correspondence study design. The matched design has several distinct advantages. First, the required experimental units to reach satisfying levels of power is halved (or further reduced when more than two applicants are submitted) for matched designs. This is an important advantage, especially when studying smaller labor markets or specific occupations, and especially when the degree of concordance is high – as is the case in many experiments on hiring discrimination. Second, the matched design arguably produces more convincing evidence of direct discrimination, which can be a particular strength when planning correspondence studies in collaboration with community groups and other organizations. Whether or not the designs substantially measure different things is a discussion that is likely to continue in the future, but outside of scholarly interest, matched designs can produce measures that are both easier to communicate and mobilize for enforcement purposes. Finally, matched designs provide direct measures on more subtle aspects of the hiring process, such as the order in which callbacks are given and the qualitative differences in responses.

The unmatched designs, apart from being robust towards the sensitivity to applicant pool sizes I describe, also has a several distinct advantages. First, although possible in matched designs, unmatched designs easily permits the inclusion of multiple orthogonal treatments without harming statistical power[12] . Second, the risk of detection is lower in unmatched studies, especially compared to matched studies with larger applicant sets. Related to this is the question of ethics, where unmatched designs might have an easier time gaining approval by review boards for the comparatively lesser degree of intrusion involved. Finally, unmatched designs arguably produce more realistic measurements of overall disadvantage by pitting fictitious applicants against real-world competitors with a background of "realistic noise". All of these factors needs to be taken into account when choosing a correspondence study design, and I urge researchers to reflect on and discuss the choice in future correspondence studies.

A potential solution that retains key strengths from both approaches, but that has not been utilized much in the literature so far, is to use *unmatched paired* designs. This is suggested by Neumark (2018) when discussing solutions to the spillover effect problem raised by Phillips (2016): "…the problem can be avoided by forgoing matched-pair designs – for example, randomizing race for each applicant independently within pairs – so that the "treatment" of being assigned a black name is not correlated with the applicant pool composition (as affected by the researcher)" (Neumark, 2018, p. 824–825). Jackson (2009) employed such a design in a study of discrimination in in the UK, where the treatment was randomly allocated to each applicant independently within pairs. This kind of designs retains desirable properties of the matched approach related to sample sizes and efficiency, while alleviating the concern related to the induced competition in matched designs where treated candidates are always paired with a reference competitor. However, they still face the issues related to detection and ethics discussed in this article. Nonetheless, I see this approach as a fruitful direction for correspondence studies and one that warrants more attention, especially in small labor markets with limited access to experimental units.

Finally, even though my argument has been formulated with regards to correspondence studies on ethnic hiring discrimination, my line of reasoning extends to audit studies, and studies of discrimination in other contexts and along other dimensions than race or ethnicity. As attention is moving beyond simply asserting the existence of discrimination in various arenas, towards attempting to explain differences and variation in its prevalence, the need for well-founded methodological decisions becomes even more crucial.

Acknowledgements

I am grateful for help and comments from S. Michael Gaddis, Torkild H. Lyngstad and Gunn E. Birkelund, as well as the two anonymous reviewers.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[12] As long as any interaction terms of interest are taken into account in power analysis: see Larsen and Di Stasio (2019) for a discussion and an example of a post-hoc power analysis for an unmatched, multi-treatment correspondence study.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.rssm.2020.100475.

References

Agan, A., & Starr, S. B. (2016). *Ban the box, criminal records, and statistical discrimination: A field experiment (pp. 16–012)*. University of Michigan Law School, Law and Economics Research Paper Series.

Agerström, J., Björklund, F., Carlsson, R., & Rooth, D. O. (2012). Warm and competent Hassan= cold and incompetent Eric: A harsh equation of real-life hiring discrimination. *Basic and Applied Social Psychology, 34*(4), 359–366.

Andriessen, I., Nievers, E., Dagevos, J., & Faulk, L. (2012). Ethnic discrimination in the Dutch Labor Market: Its relationship with job characteristics and multiple group membership. *Work and Occupations, 39*, 237–239.

Arceo-Gomez, E. O., & Campos-Vazquez, R. M. (2014). Race and marriage in the labor market: A discrimination correspondence study in a developing country. *The American Economic Review, 104*, 376–380.

Baert, S. (2018). Hiring discrimination: An overview of (almost) all correspondence experiments since 2005. In S. M. Gaddis (Ed.). *Audit studies: Behind the scenes with theory, method, and nuance* (pp. 63–77). Cham: Springer.

Baert, S., & Vujić, S. (2016). Immigrant volunteering: A way out of labour market discrimination? *Economics Letters, 146*, 95–98.

Baert, S., Cockx, B., Gheyle, N., & Vandamme, C. (2015). Is there less discrimination in occupations where recruitment is difficult? *ILR Review, 68*(3), 467–500.

Baert, S., Albanese, A., du Gardein, S., Ovaere, J., & Stappers, J. (2017). Does work experience mitigate discrimination? *Economics Letters, 155*, 35–38.

Bendick, M. (1996). *Discrimination against racial-ethnic minorities in access to employment in the United States.* Empirical Findings from Situation Testing: Employment Department, International Labour Office.

Bendick, M., Jackson, C. W., Reinoso, V. A., & Hodges, L. E. (1991). Discrimination against Latino job applicants: A controlled experiment. *Human Resource Management, 30*(4), 469–484.

Bendick, M., Rodriguez, R. E., & Jayaraman, S. (2010). Employment discrimination in upscale restaurants: Evidence from matched pair testing. *The Social Science Journal, 47*(4), 802–818. https://doi.org/10.1016/j.soscij.2010.04.001.

Berson, B. (2012). *Does competition induce hiring equity? Documents de travail du Centre d'Economie de la Sorbonne 12019*.

Birkelund, G. E., Heggebø, K., & Rogstad, J. (2017). Additive or multiplicative disadvantage? The scarring effects of unemployment for ethnic minorities. *European Sociological Review, 33*(1), 17–29.

Blank, R., Dabady, M., & Citro, C. (2004). *Measuring racial discrimination.* Washington, DC: National Academy Press.

Blommaert, L., Coenders, M., & van Tubergen, F. (2014). Discrimination of Arabic named applicants in the Netherlands: An internet-based field experiment examining different phases in online recruitment procedures. *Social Forces, 92*, 957–982.

Bonoli, G., & Fossati, F. (2018). More than noise? Explaining instances of minority preference in correspondence studies of recruitment. *Journal of Ethnic and Migration Studies,* 1–17.

Booth, A. L., Leigh, A., & Varganova, E. (2012). Does ethnic discrimination vary across minority groups? Evidence from a field experiment. *Oxford Bulletin of Economics and Statistics, 74*, 547–573.

Bovenkerk, F. (1992). *Testing discrimination in natural experiments: A manual for international comparative research on discrimination on the grounds of "race" and ethnic origin.* Geneva: International Labour Office.

Bursell, M. (2014). The multiple burdens of foreign-named men—Evidence from a field experiment on gendered ethnic hiring discrimination in Sweden. *European Sociological Review, 30*, 399–409.

Capéau, B., Eeman, L., Groenez, S., & Lamberts, M. (2012). *Two concepts of discrimination: Inequality of opportunity versus unequal treatment of equals. Ecore Discussion Papers.* 58.

Carlsson, M., & Rooth, D. O. (2012). Revealing taste-based discrimination in hiring: A correspondence testing experiment with geographic variation. *Applied Economics Letters, 19*, 1861–1864.

Carlsson, M., Fumarco, L., & Rooth, D. O. (2015). Does the design of correspondence studies influence the measurement of discrimination? *IZA Journal of Migration, 3*(1), 11.

Carlsson, M., Fumarco, L., & Rooth, D. O. (2018). Ethnic discrimination in hiring, labour market tightness and the business cycle-evidence from field experiments. *Applied Economics, 50*(24), 2652–2663.

Cherry, F., & Bendick, M. (2018). Making it count: Discrimination auditing and the activist scholar tradition. In S. M. Gaddis (Ed.). *Audit studies: Behind the scenes with theory, method, and nuance* (pp. 45–62). Cham: Springer.

Darolia, R., Koedel, C., Martorell, P., Wilson, K., & Perez-Arce, F. (2016). Race and gender effects on employer interest in job applicants: New evidence from a resume field experiment. *Applied Economics Letters, 23*, 853–856.

Decker, S. H., Ortiz, N., Cassia, S., & Hedberg, E. (2015). Criminal stigma, race, and ethnicity: The consequences of imprisonment for employment. *Journal of Criminal Justice, 43*, 108–121.

Derous, E., Ryan, A. M., & Nguyen, H. H. (2012). Multiple categorization in resume screening: Examining effects on hiring discrimination against Arab applicants in field and lab settings. *Journal of Organizational Behavior, 33*, 544–570.

Drydakis, N. (2012). Estimating ethnic discrimination in the labour market using experimental data. *Southeast European and Black Sea Studies, 12*, 335–355.

Duguet, E., Leandri, N., L'Horty, Y., & Petit, P. (2010). Are young french jobseekers of ethnic immigrant origin discriminated against? A controlled experiment in the Paris Area. *Annals of Economics and Statistics,* (99/100), 187–215.

Edo, A., Jacquemet, N., & Yannelis, C. (2013). *Language skills and homophilous hiring discrimination: Evidence from gender- and racially-differentiated applications.* CES Working Paper Series13–58.

Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*(4), 1451–1479.

Gaddis, S. M. (2018). An introduction to audit studies in the social sciences. In S. M. Gaddis (Ed.). *Audit studies: Behind the scenes with theory, method, and nuance* (pp. 3–44). Cham: Springer.

Galarza, F. B., & Yamada, G. (2014). Labor market discrimination in Lima, peru: Evidence from a field experiment. *World Development, 58*, 83–94.

Ghumman, S., & Ryan, A. M. (2013). Not welcome here: Discrimination towards women who wear the Muslim headscarf. *Human Relations, 66*(5), 671–698.

Heckman, J. J. (1998). Detecting discrimination. *The Journal of Economic Perspectives,* 101116.

Heckman, J. J., & Siegelman, P. (1993). The Urban institute audit studies: Their methods and findings. In M. Fix, & R. Struyk (Eds.). *Clear and convincing evidence: Measurements of discrimination in America* (pp. 187–258). Washington, DC: The Urban institute Press.

Jackson, M. (2009). Disadvantaged through discrimination? The role of employers in social stratification. *The British Journal of Sociology, 60*(4), 669–692.

Jacquemet, N., & Yannelis, C. (2012). Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market. *Labour Economics, 19*, 824–832.

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

Kaas, L., & Manger, C. (2012). Ethnic discrimination in Germany's labour market: A field experiment. *German Economic Review, 13*, 1–20.

Kang, S. K., DeCelles, K. A., Tilcsik, A., & Jun, S. (2016). Whitened Résumés: Race and self-presentation in the labor market. *Administrative Science Quarterly, 61*(3), 469–502. https://doi.org/10.1177/0001839216639577.

Koopmans, R., Veit, S., & Yemane, R. (2018). *Ethnische Hierarchien in der Bewerberauswahl: Ein Feldexperiment zu den Ursachen von Arbeitsmarktdiskriminierung. WZB Discussion Paper* SP VI 2018-104. Available at:https://bibliothek.wzb.eu/pdf/2018/vi18-104.pdf.

Lahey, J., & Beasley, R. (2018). Technical aspects of correspondence studies. In S. M. Gaddis (Ed.). *Audit studies: Behind the scenes with theory, method, and nuance* (pp. 81–101). Cham: Springer.

Lancee, B., Birkelund, G., Coenders, M., Di Stasio, V., Fernández Reino, M., Heath, A., Koopmans, R., et al. (2019a). *The GEMM Study: A Cross-National Harmonized Field Experiment on Labour Market Discrimination – Codebook.* Retrieved from http://gemm2020.eu/wp-content/uploads/2019/02/GEMM-WP3-codebook.pdf.

Lancee, B., Birkelund, G., Coenders, M., Di Stasio, V., Fernández Reino, M., Heath, A., Koopmans, R., et al. (2019b). *The GEMM Study: A Cross-National Harmonized Field Experiment on Labour Market Discrimination – Technical Report* Retrieved from http://gemm2020.eu/wp-content/uploads/2019/02/GEMM-WP3-technical-report.pdf.

Larsen, E. N., & Di Stasio, V. (2019). Pakistani in the UK and Norway: different contexts, similar disadvantage. Results from a comparative field experiment on hiring discrimination. *Journal of Ethnic and Migration Studies*. https://doi.org/10.1080/1369183X.2019.1622777.

Lee, H. A., & Khalid, M. A. (2016). Discrimination of high degrees: Race and graduate hiring in Malaysia. *Journal of the Asia Pacific Economy, 21*, 53–76.

Maurer-Fazio, M. (2012). Ethnic discrimination in China's internet job board labor market. *IZA Journal of Migration, 1*, 1–24.

McGinnity, F., & Lunn, P. D. (2011). Measuring discrimination facing ethnic minority job applicants: An Irish experiment. *Work Employment & Society, 25*, 693–708.

Midtbøen, A. H. (2013). *Determining discrimination. A multi-method study of employment discrimination among descendants of immigrants in Norway.* PhD dissertationUniversity of Oslo.

Midtbøen, A. H. (2016). Discrimination of the second generation: Evidence from a field experiment in Norway. *Journal of International Migration and Integration, 17*, 253–272.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature, 56*(3), 799–866.

Nunley, J. M., Pugh, A., Romero, N., & Seals, R. A. (2015). Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment. *The BE Journal of Economic Analysis & Policy, 15*, 1093–1125.

Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal Economic Policy, 3*, 148–171.

Pager, D. (2007). The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. *The Annals of the American Academy of Political and Social Science, 609*, 104–133. https://doi.org/10.2307/25097877.

Pager, D. (2016). Are firms that discriminate more likely to go out of business? *Sociological Science, 3*, 849–859.

Pager, D., Bonikowski, B., & Western, B. (2009). Discrimination in a low-wage labor market: A field experiment. *American Sociological Review, 74*(5), 777–799. https://doi.org/10.2307/27736094.

Pager, D., & Western, B. (2012). Identifying discrimination at work: The use of field experiments. *The Journal of Social Issues, 68*(2), 221–227.

Phillips, D. C. (2016). Do comparisons of fictional applicants measure discrimination when search externalities are present? Evidence from existing experiments. *The Economic Journal*.

Pierné, G. (2013). Hiring discrimination based on national origin and religious closeness:

Results from a field experiment in the Paris area. *IZA Journal of Labor Economics, 2*, 4.
Riach, P. A., & Rich, J. (2004b). Deceptive field experiments of discrimination: are they ethical? *Kyklos, 57*(3), 457–470.
Riach, P. A., & Rich, J. (2004a). Fishing for discrimination. *Review of Social Economy, 62*(4), 465–486. https://doi.org/10.1080/0034676042000296227.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.
Vuolo, M., Uggen, C., & Lageson, S. (2016). Statistical power in experimental audit studies: Cautions and calculations for matched tests with nominal outcomes. *Sociological Methods & Research, 45*(2), 260–303.
Vuolo, M., Uggen, C., & Lageson, S. (2017). Race, recession, and social closure in the low wage labor market: Experimental and observational evidence. *Research in the Sociology of Work, 30*, 141–183.
Vuolo, M., Uggen, C., & Lageson, S. (2018). To match or not to match? Statistical and substantive considerations in audit design and analysis. In S. M. Gaddis (Ed.). *Audit studies: Behind the scenes with theory, method, and nuance* (pp. 119–140). Cham: Springer.
Weichselbaumer, D. (2015). Testing for discrimination against lesbians of different marital status: A field experiment. *Industrial Relations: A Journal of Economy and Society, 54*(1), 131–161.
Weichselbaumer, D. (2016). *Discrimination against female migrants wearing headscarves. IZA Discussion Paper Series, 10217*.
Zschirnt, E. (2019). Research ethics in correspondence testing: An update. *Research Ethics, 15*(2), 1–21. https://doi.org/10.1177/1747016118820497.
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*(7), 1115–1134.