Research in Social Stratification and Mobility 65 (2020) 100463

Contents lists available at ScienceDirect

Research in Social Stratification and Mobility

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

Understanding why employers discriminate, where and against whom: The potential of cross-national, factorial and multi-group field experiments

Valentina Di Stasio[a,*], Bram Lancee[b]

[a] Utrecht University, the Netherlands
[b] University of Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Employers
Hiring
Factorial
Migration
Field experiments
Ethnic discrimination
Comparative research

ABSTRACT

While the field of (experimental) discrimination research is rapidly expanding and technology decreases the costs of designing and conducting field experiments, virtually all existing studies refer to a single country. Yet, cross-national comparison is a cornerstone of stratification and inequality research and comparative designs are necessary to understand the conditions under which employers are more prone to making biased decisions. Furthermore, previous studies often include only a handful of -typically the most marginalized- minority groups and restrict the experimental variation to ethnicity. We argue for a research design that reflects the geographical and demographic complexity of contemporary societies and is better suited to test the theoretical assumptions underlying discrimination. We discuss how external validity can be maximized in a comparative research design that is both factorial (simultaneously varying multiple treatments) and double-comparative (comparing multiple origin groups in multiple destination countries). Drawing on our first-hand experience with the GEMM study, a cross-nationally harmonized field experiment on ethnic discrimination in hiring, we show how this design can offer researchers new insight into the targets, drivers and scope conditions of employers' discriminatory behavior.

1. Introduction

Field experiments are based on the random assignment of participants to manipulated conditions and take place in natural settings. Participants engage in genuine tasks and, because they are unaware of being involved in a study, their behavior is unaffected by social desirability concerns (Baldassarri & Abascal, 2017; King, Hebl, Botsford Morgan, & Ahmad, 2013). The possibility to retain experimental control outside of the laboratory and isolate causal effects in realistic settings, is probably the most appealing feature of field experiments (Pager, 2007; Pager & Shepherd, 2008). When the focus is on hiring discrimination, these studies typically compare the interview invitations or callbacks that fictitious job applicants identical in all respects except for the alleged cause of discrimination (e.g. ethnicity, race, or gender) receive from employers.

Over the years, an increasing number of field experiments have provided compelling evidence of ethnic discrimination in the labour market, capitalizing on the opportunity to attain high external validity without a loss of internal validity (Bertrand & Duflo, 2017; Bertrand & Mullainathan, 2004; Kaas & Manger, 2012; Oreopoulos, 2011; Pager, Western, & Bonikowski, 2009). In particular, correspondence tests based on written applications can create 'blind testers' at a much lower costs than in-person audits and have come to dominate the field of experimental research on discrimination (Neumark, 2018). Taking stock of the literature, a recent meta-analysis of correspondence tests conducted in OECD countries between 1990 and 2015 found that, on average, to be invited for a job interview, minority applicants have to write fifty per cent more applications than the majority group (Zschirnt & Ruedin, 2016). Though field experiments have considerably expanded our knowledge of the discrimination experienced by ethnic minorities, important questions in discrimination research have remained unanswered. In this paper, we discuss how methodological innovations in the design of field experiments, such as the use of cross-national, factorial and multi-group designs, can help us to better understand employers' discriminatory behavior, its scope conditions, drivers and targets.

First, besides extensive discrimination, the meta-analysis by Zschirnt and Ruedin (2016) also documented considerable variation across countries. Similarly, another meta-analysis found ubiquitous discrimination towards non-white minorities of African, Middle Eastern or Asian descent, irrespective of their national origin, but large variation in the levels of discrimination experienced by these groups across countries of destination (Quillian et al., 2019). These contributions point to the need for research that links discriminatory outcomes to the context of employment. Already

* Corresponding author at: Padualaan 14, Sjoerd Groenman building, B2.06, 3584CH Utrecht, the Netherlands.
E-mail address: v.distasio@uu.nl (V. Di Stasio).

https://doi.org/10.1016/j.rssm.2019.100463
Received 5 April 2019; Received in revised form 18 December 2019; Accepted 19 December 2019
Available online 24 December 2019

more than a decade ago, Pager (2007: p.120) concluded a methodological article on the advantages of using experimental field methods for studying discrimination with the following call:

> "*it would be useful for future research to develop a standardized audit framework that could be replicated across testing sites and over time… Though several researchers have conducted multicity studies, no researcher has attempted to include more than two sites, thus limiting our comparative perspective on discrimination across labor markets and over time.*"

From a theoretical point of view, comparative research designs can generate new insights into the contexts *where* discrimination is more or less likely to occur, thus shedding light on its scope conditions. For example, the 'new institutionalist' theory emphasizes how employers' recruitment and rewarding behaviour depends on the institutional context in which they make their decisions (Brinton & Nee, 1998; Midtbøen, 2015). From a more practical point of view, the possibility to automate the creation of randomly varying resumes and the increasing popularity of online job posting sites have exponentially decreased the time and costs of fieldwork (Carbonaro & Schwarz, 2018; Lahey & Beasley, 2009). Hence, comparative field experiments are not only desirable, but have also become increasingly feasible.

Second, over the past two decades societies have become so increasingly diverse that some scholars have coined the term 'superdiversity' to reflect the heterogeneity and substantial size of ethnic groups currently living in large Western European cities (Crul, 2016; Vertovec, 2007). Double comparative designs, simultaneously varying both the origin country and the destination country of the job applicant, are a promising way forward to analyse and compare the discrimination experienced by different ethnic origin groups in multiple destination countries. Given the possibility to automate the search for job vacancies and the generation of bogus applications, the number of groups can be large. Furthermore, designs that are factorial, allow for the randomization of other characteristics in addition to ethnicity, and can thus mirror the geographical and demographic complexity of contemporary migration flows. As such, factorial double comparative designs are ideal to understand *who* is being discriminated against.

A third longstanding debate within the discrimination literature is *why* ethnic minorities are discriminated against. The direct manipulation of the expected drivers of discrimination in the study design is one of the emerging frontiers in field experimental research on this topic (Pedulla, 2018). Factorial designs can accommodate multiple experimental variables into a single study, allowing researchers "to calibrate the effects of race against other key labor market determinants" (Pager, 2007, p. 120) and to disentangle the separate contribution of characteristics that typically co-occur in the real world. For example, to distinguish between ethnic origin discrimination and discrimination based on religious grounds, religion and ethnicity can be varied independently (Di Stasio, Lancee, Veit, & Yemane, 2019; Koopmans, Veit, & Yemane, 2019). To test whether the generally negative portrayal of Muslim men in public debates and in the media also affects employers' perceptions, gender can be added to the design. In an effort to distinguish statistical from taste-based discrimination, skills and other information relevant to assess job candidates can be varied too.

Although field experimental studies on ethnic discrimination are plenty, virtually all existing work consists of single-country studies. Moreover, most research focuses on the most marginalized groups in society; a design choice that may overstate labor market discrimination (Gaddis, 2019; but see Bessudnov & Shcherbak, 2019, and Koopmans et al., 2019 for noteworthy exceptions). In the *Growth, Equal opportunities, Migration and Markets* (GEMM) study,[1] we paid heed to the call for comparative and factorial designs. The GEMM study is a collaborative effort involving five teams of researchers in five European countries (Britain, Germany, the Netherlands, Norway and Spain) with the explicit aim to conduct a cross-nationally harmonized correspondence test in five different institutional contexts (for an overview of the data, see Lancee, 2019). The design of the GEMM study is cross-national, multi-group and factorial: next to ethnicity, the bogus applicants vary in gender, religious affiliation, grades, country of birth (foreign-born vs. 2nd generation migrants) and self-reported information on job-related competences and social skills. A total of 53 origin groups were included, ranging from Europe to the African continent, to Asia and the Middle East, and varying in language, culture, colonial ties with the host countries, and religion.

With this double-comparative design (cross-national and multi-group), the GEMM study is well-suited to analyze the labor market integration of multiple ethnic minorities in multiple destination countries. Furthermore, the large number of origin groups and the factorial design containing several treatment conditions increase the external validity of the study and enable comparison of discrimination rates across different sub-group targets. In what follows, we address the issues that researchers face while conducting comparative research on ethnic and racial discrimination using field experiments, with a particular focus on their external validity. We then take the GEMM study as an example of how methodological choices regarding the experimental design of correspondence tests can provide researchers with the opportunity to address pressing questions in the field of discrimination research. We also discuss the challenges we encountered while planning and implementing the study and the trade-offs we were confronted with when deciding on the experimental design. Finally, we show the implications of using a double-comparative and factorial design on the findings, with a focus on the conclusions that can be drawn about differences in discrimination across contexts and groups.

To foreshadow some of our findings, in our analysis we show that "where" matters: discrimination varies substantially across countries, with the highest level of discrimination in the United Kingdom and Norway, and the lowest level in Germany. However, even more relevant is the question 'who' is discriminated: we observe the largest variation in discrimination *across* ethnic minority groups, with black minorities and minorities from MENAP countries facing the highest level of discrimination, and white minorities the lowest. Last, we show that factorial designs can contribute to answering the 'why' question: signalling affiliation with Islam on the CV significantly reduces callback rates over and above one's country of origin.

2. Research design issues in discrimination research

Field experiments are ever more popular in discrimination research. The experimental control over possible confounders and the experimental realism achieved with randomization in naturally occurring settings have earned field experiments the much-coveted title of 'sterling-gold standard' of organizational research methods, as they offer researchers the opportunity to 'grab the fabled validity stick by both ends' (Eden, 2017, p. 96).

Despite their obvious merits, field experimental designs face threats to both internal and external validity. With regard to internal validity, field experiments rely on the random assignment of ethnic cues to fictitious job applications, which yields unbiased estimates of the unfair treatment received by minority applicants in their search for a job. Critics have pointed to the possibility that experimenter effects or inadequate matching may pose a threat to internal validity, concerns that apply to a far lesser degree to correspondence tests, which rely on written applications (for a discussion: Jackson & Cox, 2013; Neumark, 2018; Pager, 2007). Lack of control over treatment conditions may still be a problem, though, insofar as the treatment used to signal group membership is not perceived by the employer as intended. In this paper, we focus on external validity. For methodological discussions on how to improve internal validity, we refer the reader to recent studies on the importance of pretesting perceptions of the names used in field experiments to ensure the construct validity of the treatment (e.g. Gaddis, 2017).

With regard to external validity, the big advantage of field experiments is that they are not confined to the artificiality of the laboratory as they take

[1] The data collection has been supported by funding from the European Commission (Grant number H2020 649255).

place 'in the wild' (Jackson & Cox, 2013). However, pursuing randomization in naturally occurring settings is not sufficient for results to be externally valid. A first issue is the limited potential for generalization across labor market segments. In correspondence tests on hiring discrimination, most researchers readily acknowledge that the range of sampled job openings is usually limited to white-collar occupations, often in female-dominated sectors such as administration, sales or customer service, which reduces the external validity of the findings[2] . Furthermore, given that nearly all existing studies are single-country designs, cross-national generalization is even more problematic. Second, external validity is affected by stimulus sampling, i.e. the sample of experimental units used in the study (Highhouse, 2009; Jackson & Cox, 2013), an issue largely overlooked in research on discrimination. For example, only a handful of minority groups are usually included in the design of audit and correspondence tests. Typically, these are the most marginalized groups in society. An additional threat to external validity comes from the fact that the matched pairs design requires holding constant all characteristics of the applicants except for their ethnic or racial background, thus drastically minimizing the demographic complexity of minority groups. Below, we address three methodological advancements in the design of correspondence tests that can address these concerns about external validity, as well as bridge the research gaps identified in the *introduction*.

2.1. Innovation I: cross-nationally harmonized field experiments

Previous field experiments on ethnic discrimination have either focused on a single origin group in a single destination country (as, for example, in the seminal study by Bertrand & Mullainathan, 2004) or, when multiple origin groups were compared, the destination country was held constant (Andriessen et al., 2012; Booth, Leigh, & Varganova, 2012; Weichselbaumer, 2017; Wood, Hales, Purdon, Sejersen, & Hayllar, 2009).

Research that compares discrimination rates across multiple destination countries is scant. One often-cited study, initiated by the ILO, compared ethnic discrimination in Germany, Spain, the Netherlands, and Belgium (Bovenkerk, 1992; Zegers de Beijl, 2000). However, due to cross-national differences in the volume and type of job openings available, the standardization of the experimental protocols proved difficult. Moreover, the guideline to include, in each country, two large migrant and ethnic minority groups was a rather loose standard for harmonizing the design of the job applications across countries. Thus, any cross-national difference in the discrimination rates could be due to the different minority groups or types of job targeted in the participating countries[3] .

Given the lack of comparative field experiments on ethnic discrimination, an emerging line of cross-national research is based on meta-analysis. In a meta-analysis, the discrimination ratios reported for different minority groups or countries in single studies are summarized in a pooled estimate. While this approach has some appeal (Ross, 2017), a disadvantage of meta-analyses is that they lack a truly comparative design: any difference in the effect sizes could in principle be an artefact of the specific design adopted in each single study (Pedulla, 2018)[4] .

The benefits of a cross-national design are twofold. First, it offers a descriptive and comparable account of the level of discrimination experienced by ethnic and racial minorities in different countries. This is important, for example, to inform policy makers on the severity of discrimination at the societal level, benchmarking it against that of other countries. Second, a big asset of cross-national designs is that they allow for a comparison of institutional contexts. For example, an often-theorized but seldomly tested hypothesis is that the flexibility of the labour market affects employer's discriminatory decisions (Kogan, 2006; Lancee, 2016; Larsen & Di Stasio, 2019). Similarly, the argument that in German-speaking countries the more formalized application procedures may explain lower levels of discrimination, often labelled the 'German exceptionalism' (Zschirnt & Ruedin, 2016), can only be tested with cross-national designs comparing Germany with countries that have less strict application standards.

2.2. Innovation II: double-comparative field experiments

As migration flows are often the historical by-product of post-colonial ties or post-war industrial policies, the national origins of the largest migrant groups vary a great deal across countries of destination. For example, while a large number of migrants from Turkey, Morocco and Southern Europe moved to Germany or the Netherlands as part of guest workers recruitment programs, former South Asian colonies were the primary source of manpower to cope with labour shortages in Britain (Heath, Rothon, & Kilpi, 2008). As a result, differences in the composition of the ethnic landscape of destination countries may represent an obstacle to the harmonization of cross-national field experiments.

A promising design to study discrimination cross-nationally while retaining the interest of each participating country in origin groups that do not necessarily overlap is the double comparative design. Double comparative designs combine the strengths of *single-destination multi-group* designs comparing multiple origin groups within a single destination country with the strengths of *single-origin-multiple-destination* designs comparing the integration outcomes of the same origin group across multiple contexts (Van Tubergen, Maas, & Flap, 2004). While the former offer analytical leverage to examine how characteristics of the origin country may affect the labour market integration of immigrants, the cross-national focus of the latter is better suited to single out the role of destination country characteristics in facilitating or hindering immigrants' integration. In double comparative designs, "multiple origins in multiple destinations are compared, suggesting that the economic status of immigrants may be affected by the country from which they come ('origin effect'), the country to which they migrate ('destination effect'), and the specific relations between origins and destinations ('community effect')" (Van Tubergen et al., 2004: 704).

The inclusion of multiple groups has several advantages. First, the groups included can reflect the actual ethnic minority population, improving external validity. Second, a wide spectrum of ethnicities makes it possible to test whether employers' preferences are simply driven by ingroup favouritism and ethnic homophily (Edo, Jacquemet, & Yannelis, 2019; Jacquemet & Yannelis, 2012) or gradually patterned according to a well-defined ethnic hierarchy (Hagendoorn, 1995; Koopmans, Veit, & Yemane, 2018). Moreover, the inclusion of origin groups that vary in religion and skin colour can reveal whether ethnic hierarchies can be explained by racism, or Islamophobia or stereotype-based explanations of discrimination. For example, Koopmans et al. (2019) show German employers no longer discriminate against applicants of non-German ethnic origin, non-white phenotype and non-Christian religion once their value distance relative to the German population is taken into account. Third, multi-group designs that are also cross-national (i.e. double-comparative designs) offer researchers the opportunity to test whether employers' hiring preferences are inversely proportional to the geographical and cultural distance that separates origin groups from their countries of destination.

In double-comparative designs, researchers can prioritize *direct*

[2] A similar remark can be made on in-person audits, which are often concentrated in entry-level jobs in the low-skill sector.

[3] The only other cross-national field experiments we know of deal with other discrimination grounds, namely age (Riach, 2015) and parental status (Becker, Fernandes, & Weichselbaumer, 2019).

[4] Meta-analyses try to account for differences in design, focus and time of fieldwork across studies with an extensive list of controls that typically include the characteristics of the application (in-person vs. written), applicant (gender, level of education) or market under study (blue-collar vs. white-collar occupations, time of fieldwork). Just like in any regression analysis, however, effectively accounting for all differences across studies is nearly impossible. For example, neither Zschirnt and Ruedin (2016) nor Quillian et al. (2019) control for differences in signal strength, that is, how the ethnic minority background of the candidate is signalled in the job application.

equivalence or *functional equivalence* (Johnson, 1998). Direct equivalence refers to units that can be directly compared across contexts: for example, the same non-European origin group is analyzed in different countries, regardless of cross-national differences in group size. By contrast, functional equivalence guarantees that units are nominally different but functionally comparable or, in other words, "universal in a qualitative, although not necessarily a quantitative sense" (Van De Vijver & Poortinga, 1982). For example, the largest non-European group is sampled in each country, regardless of cross-national differences in origin country. While direct equivalence gives analytical leverage to separate contextual from group-specific explanations, functional equivalence might be preferable for testing social psychological theories of threat, which stress the role of outgroup size: functional equivalence accounts for the fact that the national origin of the largest outgroups may differ across countries. Direct and functional equivalence are often discussed as a trade-off. However, this is not necessarily the case. In the data section, we describe how the ethnic groups included in the GEMM study are a combination of directly and functionally equivalent groups.

2.3. Innovation III: factorial field experiments

Traditionally, correspondence tests have employed matched pairs designs with the aim to file litigation against biased employers (Baldassarri & Abascal, 2017). Each employer receives a pair or a set of applications that are equivalent in all respects except for the treatment of interest (e.g. foreign-sounding names or country of origin), which is randomly assigned to templates created beforehand by the researcher. Problematically, with these designs, "all items except the variable of interest are correlated within each pair of templates, so the results can only predict the outcomes and interaction effects for specific bundles of characteristics rather than individual characteristics" (Lahey & Beasley, 2009, p. 88). More generally, regardless of whether paired or unpaired, field experiments that only vary ethnicity can only uncover the discrimination experienced by applicants who are similar to the ones that the template represents. By contrast, factorial designs can include several other treatment variables in addition to ethnicity, such as gender, academic achievement and religious affiliation. Moreover, by design, the effect of ethnicity on callback can be estimated independently of all other characteristics included in the study, allowing for generalization of the ethnicity effect to a much larger number of cases than it would be possible when using only a few templates.

Factorial designs have two additional advantages. First, the inclusion of other variables in the design enables researchers to benchmark the possible stigma of belonging to an ethnic minority group against other key predictors of labor market success or disadvantage (Pager, 2007). Recent field experiments have, for example, varied both ethnicity and gender to test whether minority women are especially at a disadvantage (Bursell, 2014; Dahl & Krog, 2018). This is important in light of current debates on intersectionality and the distinctive disadvantages faced by members of multiple subordinate groups (Crenshaw, 1991; Mccall, 2008). At a time when "the ethnic group as a unit of analysis has become problematic" (Crul, 2016, p. 66) factorial designs offer the possibility to study heterogeneous effects of minority status on callbacks by education, religion, generation, socioeconomic status, etc.

Second, additional treatment variables allow for manipulating key theoretical mechanisms to better understand the drivers of discrimination (Pedulla, 2018). While the ethnicity variable provides convincing evidence regarding the question 'who' is discriminated against, additional treatment variables help us understand why discrimination occurs. For example, in an effort to distinguish between statistical and taste-based discrimination, previous studies have varied the amount or quality of information provided in the job applications (Bertrand & Mullainathan, 2004; Kaas & Manger, 2012). If employers discriminate because of information asymmetries, the gap in callbacks should reduce for applicants that provide more information in their job application (but see Neumark, 2018, p. 839 for a critical perspective on this approach). In another study, Agerström, Björklund, Carlsson, and Rooth (2012) randomly assigned signals of warmth and competence to the applicants' cover letters to analyze whether discrimination of Arab applicants was driven by employers' perceptions of this group as incompetent and lacking social skills.

3. The GEMM study: a cross-nationally harmonized field experiment with a double-comparative and factorial design

The GEMM study was carried out in five countries: Germany, The Netherlands, Britain, Spain and Norway. In these countries, a total of 19,181 fictitious applications were sent to real vacancies (for a detailed description of the data, see Lancee, 2019; Lancee, Birkelund, Coenders, Di Stasio, Fernandez Reino, Heath, Koopmans, Larsen, Polavieja, Ramos, Soiné et al., 2019, Lancee, Birkelund, Coenders, Di Stasio, Fernandez Reino, Heath, Koopmans, Larsen, Polavieja, Ramos, Thijssen et al., 2019).

The GEMM study relies on an unpaired design: only one application was sent to each single employer. Just as well as the paired design, the random allocation of treatments and controls to experimental units in the unpaired design ensures unbiased estimates, provided that the randomization process is done properly (Vuolo, Uggen, & Lageson, 2018). Applicants' age ranged from 22 to 26, depending on the occupation. All applicants were fulltime employed with four years of uninterrupted working experience.

Selecting occupations in a cross-national study is not straightforward. Nominally equivalent occupations often have different educational requirements in different national contexts, an issue that is further exacerbated by the imperfect cross-national comparability of formal qualifications. Furthermore, occupations that are equivalent in terms of entry requirements may still differ in other respects such as gender composition, labor supply, geographical distribution, to name just a few. The channels used to advertise job openings may also be country-specific. For example, we considered analyzing nursing but gave up on this idea after realizing that most nursing jobs in the British context are advertised through a separate website administered by the National Health Service and require applicants to fill in very comprehensive and standardized intake forms. Eventually, we selected six occupations that have a high degree of comparability across countries: cooks, store assistants, payroll clerks, receptionists, software developers and sales representatives.

3.1. Treatment variables

The main variable of interest, ethnicity, was randomly assigned to the job application. The GEMM study contains 53 ethnic groups: a common core of 32 directly equivalent groups, plus 21 both directly and functionally equivalent groups that are country-specific and partly over-sampled. Table 1 lists the ethnic groups included in the study.

Besides ethnicity, we varied characteristics that relate both to the person who is the target of discrimination and the reason why employers discriminate. For higher external validity and the possibility to calibrate the effect of ethnicity against other achieved or ascribed traits we varied gender, religious affiliation, headscarf, and, in the form of a profile picture, phenotype. In order to improve our understanding of why employers discriminate, information on academic achievement as well as self-reported statements on warmth (social skills) and competence were randomly assigned. A detailed description of the treatment conditions is provided in the codebook (Lancee, Birkelund, Coenders, Di Stasio, Fernandez Reino, Heath, Koopmans, Larsen, Polavieja, Ramos, Soiné et al., 2019; Lancee, Birkelund, Coenders, Di Stasio, Fernandez Reino, Heath, Koopmans, Larsen, Polavieja, Ramos, Thijssen et al., 2019) and in Lancee (2019).

Informal norms about the type of information that is considered

Table 1
Design of the GEMM study: distribution of origin countries across destination countries.
Source: GEMM, 2019.

Country of study/ majority (25%)	First oversampled minority (12.5%)	Second oversampled minority (12.5%)	Minority groups common to all countries	Minority groups specific to country of study
Britain	Pakistan	Nigeria	Albania, Britain, Bulgaria, China, Egypt, Ethiopia, France, Germany, Greece, India, Indonesia, Iran, Iraq, Italy, Japan, Lebanon, Mexico, Morocco, Netherlands, Nigeria, Norway, Pakistan, Poland, Romania, Russia, South Korea, Spain, Turkey, Uganda, USA, Vietnam	Bangladesh, Jamaica, Ireland, Somalia
Germany	Turkey	Lebanon		Dominican Republic, Macedonia, Malaysia, South Africa
Netherlands	Morocco	Turkey		Antilles, Belgium, Macedonia, Malaysia, Suriname
Norway	Pakistan	Somalia		Bosnia and Herzegovina, Denmark, Eritrea, Lithuania, Philippines, Sweden
Spain	Morocco	Ecuador		Bosnia and Herzegovina, Catalonia, Dominican Republic, Philippines, Portugal, Ukraine

appropriate in a job application, as well as more formal requirements about what documents to include, vary from one country to the other. For example, employers in German-speaking countries require not only a CV and a cover letter but also a photograph and a copy of school certificates. Due to these stricter standards, the preparation of the field experiment is not only a more cumbersome process in countries like Germany, but also more challenging from a comparative perspective. We decided to include pictures in those countries where they are either required (Germany and Spain) or commonly used in job applications (the Netherlands). Following the same principle, we included copies of certificates in German applications, but not in the other countries. Had photographs and certificates not been included, employers may have rejected a candidate simply because the application was incomplete or considered unusual.

This being said, it is important to acknowledge that cross-national differences in application standards may affect the precision and salience of the ethnicity signal. Pictures minimize or even eliminate the risk that employers may attribute a given name to the wrong ethnicity and gender, an often-overlooked issue in field experimental research that may bias the findings. From this perspective, their inclusion increases construct validity. To make sure that ethnicity was correctly identified even in those countries, Norway and Britain, where pictures were omitted, we decided to mention, in all countries, the ethnic background of the applicant in the cover letter and to include the country of origin language in the CV as a further cue. On the one hand, it is possible that the additional information conveyed through a picture makes statistical discrimination less likely (especially because all our applicants had a professional and good-looking appearance). On the other hand, it is also possible that the photographs make ethnicity more salient, rather than less: as observed by Weichselbaumer (2017: p. 244), "discrimination may be stronger when one is confronted with a photograph of an 'ethnic other' than when the migration background is indicated only abstractly by the name". With regard to the inclusion of school certificates in job applications, the fact that German employers have more – and certified – information on applicants' academic achievements, may reduce their tendency to discriminate statistically (Weichselbaumer, 2017; Zschirnt & Ruedin, 2016). All in all, researchers should be aware that they have no way of assessing the direction of any possible bias resulting from such cross-national differences in signal strength. In general, researchers are confronted with a trade-off: manipulations should be realistic in order to avoid demand effects; at the same time, they have to be strong enough to cause an effect (Highhouse, 2009; King et al., 2013).

3.2. Dependent variable

The dependent variable consists of a dichotomous variable indicating the employer's response to the application. The literature on field experiments typically differentiates between 'callback', and 'invitation to an interview'[5] . Callback is defined as any signal of positive interest from an employer, including requests to provide additional information, while invitation is restricted to an unambiguous invitation to a job interview.

While in a single country study, the choice of the dependent variable primarily relates to the researcher's interest in a specific phase of the hiring process, this is different for cross-national analysis. As can be seen in the Fig. 1, the probability of callback versus interview varies greatly across countries. Additionally, in the righthand panel, we plot

[5] Naturally, other types of employers' responses can be analysed with field experimental data, such as, for example, the number of contact attempts, the time interval until callback, the probability of being notified in case of a negative response or the tone of the negative response (Weichselbaumer, 2017).

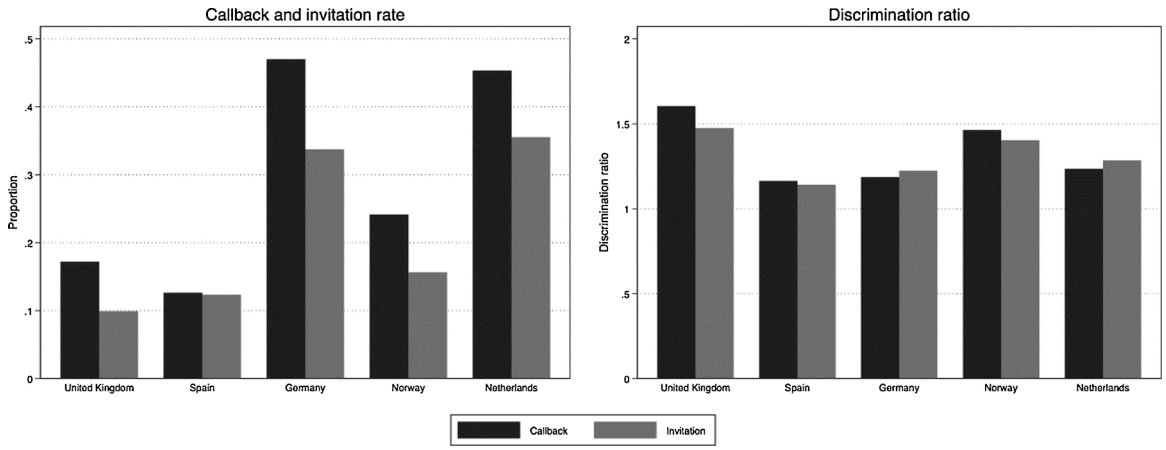

Fig. 1. Callback rate, invitation rate and discrimination ratio by country of study.
Source: GEMM, 2019.

the discrimination ratio for minority versus majority job applicants by country of study. In the UK, Norway and Spain, discrimination is lower regarding invitation than callback, while in the Netherlands and Germany this is the other way around. However, the differences are not very large.

One of the reasons for the variation in callback and invitation rate has to do with the way the application process unfolds in the different countries. For example, the share of requests for additional information as a percentage of the total number of callbacks is 3 % for Spain, but 42 % in the UK, 29 % in Germany, 21 % in Norway, and 23 % in the Netherlands. We speculate that one of the reasons why employers in the UK ask relatively more often for additional information before inviting a candidate to interview may be the vaguer language used in the job advertisements, which – coupled with the fact that British educational credentials say little about the skills people have – may mean that recruiters prefer to double-check the background of candidates before inviting them to an interview (therefore the need for an additional step in the hiring process). It is not uncommon to find sentences like 'qualified by experience' or 'formal qualifications are not essential' in British job ads. By contrast, in Germany and the Netherlands, entry qualifications are always listed in the job specification and users, when skimming through online job postings, can use the required level of education as a filter to refine their search.

Furthermore, following the guidelines of the ethics committees that gave the project IRB approval, we promptly declined any invitation or request for information. As a consequence, invitations to job interviews are rarer in the UK also because we have withdrawn the application whenever we received a request to provide additional information.

Fig. 1 thus shows that in a cross-national analysis, it is important to think about the comparison that one is interested in: comparing invitation rates follows a direct equivalence principle but ignores meaningful cross-national differences in the hiring procedure. An analysis based on callbacks captures the first step in the hiring procedure in each country (however, at the costs of ignoring possible differences in discrimination between the interview and callback phase). For the reasons outlined above, we deem the callback variable more suitable for cross-national comparison and proceed with it in the results section.

3.3. Estimation strategy

Another choice that researchers face refers to the statistical test used to establish whether a minority group is being discriminated. The most frequently used measures are the discrimination ratio, the difference in proportions, and the odds ratio (for similar discussions, see Heath & Di Stasio, 2019; Quillian et al., 2019). The measures differ in their sensitivity to the baseline callback rate, a classic problem of comparing probabilities across different marginal distributions (Goldthorpe, 2016). In single country studies, this is not an issue, as there is only one baseline callback rate. However, when callback rates differ across countries, this variation may reflect cross-national differences in the state of the economy at the time of the fieldwork, or in the extent to which the applicants used as testers are attractive relative to the available supply of labor.

The implications of using one measure instead of the other are best illustrated with an example. The difference in proportions is insensitive to the callback rate: in fictitious country A, a callback rate of 15 percent for the majority and 10 percent for the minority results in a percentage point difference of 5 %; just like in country B, with callback rates of 65 and 60 percent respectively. On the other hand, the discrimination ratio (DR, also known as likelihood ratio and calculated as $CB_{maj.} / CB_{min.}$), does vary. In country A, DR = 15/10 = 1,5; meaning that, to have the same callback chance as the majority, minority candidates have to send 50 % more applications. In country B, with 65/60 = 1,08, the DR is much lower, requiring only 8 % more applications to be at par with the majority population. The odds ratio (OR) is also sensitive to the callback rate.[6] In country A, OR = 1,59, while in country B OR = 1,39. The problem with odds ratios, however, is that their interpretation is not always straightforward. Following Polavieja, Lancee, Ramos, Veit, and Yemane (2019), we can convert the OR in discrimination ratios by using the overall callback rate of the majority[7]. In our example, $DR_{Conv.}$ = 1,28 for country A and $DR_{Conv.}$ = 1,20 for country B.

Table 2 presents these measures for the five countries included in the GEMM study. While the ordering of countries is identical across measures, the comparisons are different. The most striking difference is that between the percentage points and the percent difference: while Norway and the Netherlands both have a ten percentage-points difference in callback, due to the differences in average callback, the percentage difference and the consequential discrimination ratio is 1.46 for Norway and 1.23 for the Netherlands. The DR also differs from $DR_{Conv.}$; discrimination ratios based on odds are closer together then discrimination ratios based on percentages. In this paper, because of its ease of interpretation and its widespread use, we present findings based on the discrimination ratio (DR).

4. Analyzing double-comparative and factorial field experimental data

Above, we have discussed three advantages of cross-national, factorial and multi-group field experiments. In this section, we illustrate

[6] $OR = (CB_{maj.}/ (1 - CB_{maj.})) / (CB_{min.}/ (1 - CB_{min.}))$.

Table 2
Measures of discrimination, by country of study.
Source: GEMM, 2019.

	Percentage callback		Difference		Discrimination ratio		Odds ratio		
	Majority	Minority	%points	%	DR	[95 % CI]	OR	[95 % CI]	$DR_{Conv.}$
UK	24	15	9	60	1.60	1.37 – 1.84	1.80	1.47 – 2.19	1.41
Spain	14	12	2	16	1.16	0.95 – 1.37	1.19	0.96 – 1.47	1.12
Germany	53	45	8	19	1.19	1.09 – 1.29	1.40	1.18 – 1.66	1.23
Norway	32	22	10	46	1.46	1.25 – 1.68	1.68	1.35 – 2.08	1.36
Netherlands	53	43	10	23	1.23	1.14 – 1.33	1.50	1.29 – 1.73	1.27

Notes: $DR = CB_{maj.} / CB_{min.}$ and $OR = (CB_{maj.}/ (1 - CB_{maj.})) / (CB_{min.}/ (1 - CB_{min.}))$. We also convert OR into $DR_{Conv.}$ using the overall callback rate for the majority group across the five countries of study, using the following formula: $DR\,conv. = OR/ ((1 - overall\ CB_{maj.}) + (overall\ CB_{maj.} * OR))$. 95 % CIs for DRs are calculated using the Delta method. Sample according to functional equivalence.

Table 3
Proportion of occupations by country of study.
Source: GEMM, 2019.

	Britain	Spain	Germany	Norway	Netherlands	Total
Cook	12.2	39.0	16.6	15.0	22.4	22.2
Payroll Clerk	28.1	18.3	16.6	17.9	19.0	20.1
Receptionist	13.7	12.6	16.7	3.8	12.0	12.2
Sales Representative	17.8	5.5	16.7	30.6	16.5	16.2
Software Developer	14.1	5.5	16.7	18.8	16.7	13.8
Store Assistant	14.0	19.1	16.7	13.8	13.4	15.5
Total	100.0	100.0	100.0	100.0	100.0	100.0

how the implementation of these three methodological advancements in the GEMM field experiment allows us to address important, and still unanswered, questions on the drivers, targets and scope conditions of employers' discriminatory behavior.

4.1. 'Where' employers discriminate

When analyzing discrimination cross-nationally, one needs to think about the analytic sample. If the focus is on testing whether ethnic minorities are treated more favorably in one context compared to another, direct equivalence allows disentangling institutional explanations, such as the legislation or policies in place in the context of reception, from origin-related causes, such as group-specific stereotypes.

We proceed with an example based on direct equivalence. To directly compare discrimination ratios across all countries, we restrict the sample to the ethnic groups included in all countries, and weigh them equally. Second, we account for cross-national differences in the occupational structure. In every field experiment, besides budget and time, the sample size is a function of the amount and type of vacancies that is published. The sample thus depends on the economic cycle and the occupational structure in the country of study. Table 3 presents the proportional distribution of the occupations by country of study 'as collected'. As the propensity of employers to discriminate vary across occupations depending on, among others, formal qualification requirements and intensity of customer contact (Andriessen et al., 2012; Midtbøen, 2015), in the unweighted sample cross-national differences in discrimination may be confounded by cross-national differences in the sampled occupations. To account for these differences, we weigh all occupations equally[8].

Fig. 2 presents the discrimination ratio for the five countries of study. Discrimination is highest in Britain with a DR of 1,54, followed by Norway (DR = 1,38) and Spain (DR = 1,33). The Netherlands (DR = 1,20) and Germany (DR = 1.15) show the lowest levels of discrimination. The 'comparative' question is whether these differences can be linked to institutional explanations. Put differently: does the employment context matter? In line with the meta analytical results (Quillian et al., 2019; Zschirnt & Ruedin, 2016) and the hypothesis that discrimination is lower in contexts with a more formal and extensive application procedure, also when using harmonized experimental data, discrimination is lowest in Germany. However, in our study the discrimination ratio in Germany is not substantially different from the one in the Netherlands, a country where applications are much less formal. It is important to note that cross-national differences may still be present when focusing on specific groups (see, for example, the comparison of Turkish minorities in Thijssen, Lancee, Veit, & Yemane, 2019), which underscores the added value of double comparative designs.

Another longstanding cross-national hypothesis that we briefly addressed in the theory section is the idea that discrimination should be lower in more flexible labour markets. Yet, the highest discrimination is observed in Britain, the country in our sample with by far the most flexible labour market. Based on these five countries, this hypothesis thus does not find any support; a conclusion also drawn by Larsen and Di Stasio (2019) in a more focused comparison of the Pakistani minority in Britain and Norway. The high level of discrimination experienced by minorities in the British labor market is particularly puzzling in light of the relatively more advanced anti-discrimination legislation in place in

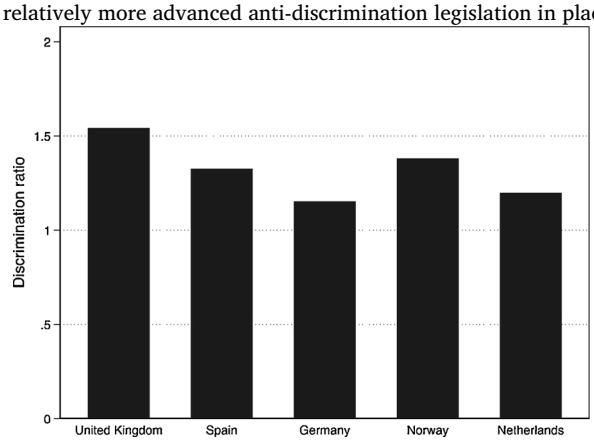

Fig. 2. Discrimination ratio, by country of study.
Note: The discrimination ratio is calculated as $DR = CB_{maj.} / CB_{min}$. *The figures are adjusted for compositional differences in ethnic origin and occupation by assigning weights.*
Source: GEMM, 2019.

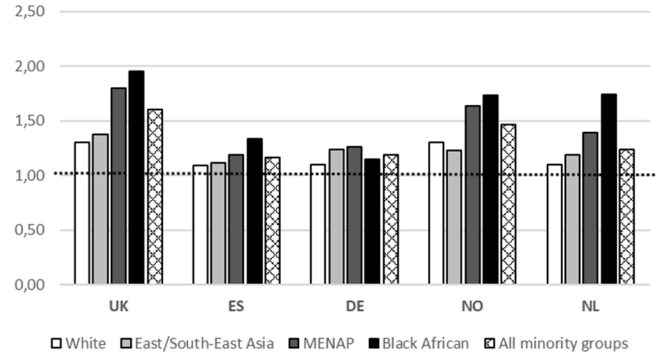

Fig. 3. Discrimination ratio by origin group and country of study.
Note: The discrimination ratio is calculated as DR = $CB_{maj.}$ / CB_{min}. Unweighted data.
Source: GEMM, 2019.

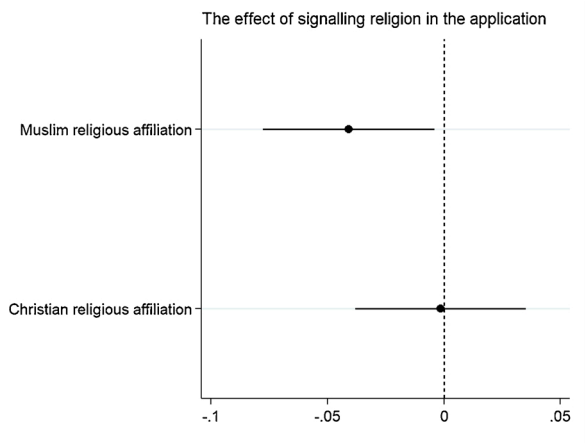

Fig. 4. Are employers discriminating against Muslim applicants because of Islamophobia?
Note: Coefficients from a linear probability model regressing callback on the religion signal, including country and occupation fixed effects. The analysis is limited to minority applicants from Albania, Bulgaria, Egypt, Ethiopia, Indonesia, Lebanon, Nigeria, Russia, Uganda (N = 2783). Minority applicants who did not mention any religious affiliation are the reference category. 95 % confidence intervals.
Source: GEMM, 2019.

Britain, compared to the other countries. These results may seem counterintuitive and cast doubt on the capacity of national institutions to limit ethnic discrimination. While it is difficult to disentangle institutional effects based on only a handful of countries (an aspect we return to in the conclusions), if anti-discrimination legislation and flexible employment protection had any role in reducing discrimination, we should still have observed a lower level of discrimination in Britain, the case that most fits these theoretical explanations (most-likely case study). The fact that we do not suggests that employers' hiring decisions are driven more by animus than by rational considerations or that national-level institutions such as anti-discrimination legislation are of little value if not properly enforced within the workplace.

4.2. 'Who' is discriminated against

While Fig. 2 clearly shows that discrimination varies across countries, it masks variation across ethnic groups. Fig. 3 differentiates callback ratios by the geographic origin of the applicant. We grouped ethnic minorities in four categories: white minorities originating from a European country, Russia or the United States; minorities from East and South-East Asia, MENAP minorities from the Middle East, North Africa or Pakistan, and Black African minorities from Sub-Saharan Africa[7] .

Glancing at Fig. 3, three observations can be made. First, in all countries but Germany, white minorities experience less discrimination than black minorities, suggesting that race is one of the drivers of employers' bias. Second, MENAP groups, who are typically the focus of correspondence studies (for example, Moroccans are often analyzed in the Netherlands and Spain, Turks in Germany, Pakistani in Norway or the UK), experience relatively high levels of discrimination compared to other minority groups. This implies that results from previous field experiments cannot be generalized to the total minority population, as doing so would overestimate the level of discrimination. Third, cross-national variation in discrimination is more pronounced when looking at culturally and geographically distant minority groups such as MENAP and black minorities. Taken together, these findings suggest that employers are not indiscriminately discriminating as suggested by earlier studies (Andriessen et al., 2012; Edo et al., 2019; Jacquemet & Yannelis, 2012). Differences between groups only become visible when the sample includes a large and varied set of ethnic minorities. Rather than by ethnic homophily, and a general reluctance to hire members of ethnic outgroups, in the GEMM study, employer's preferences can be explained by a reluctance to hire members of culturally distant groups. More specifically, when restricting the comparison to culturally distant groups, the relatively lower discrimination experienced by East Asians and South East Asians compared to the MENAP groups points to anti-Muslim attitudes as one possible driver of employers' hiring behavior.

4.3. 'Why' employers discriminate

We illustrate how a factorial design can help answering 'why' employers discriminate with an example. Because religion and ethnicity are highly correlated, it is not easy to distinguish ethnic discrimination from religious discrimination (Heath & Martin, 2013). Islam is often problematized in public debates as a religion that is incompatible with the values and orientations of Western European societies. The majority feels symbolically threatened by the religious practices of Muslims and avoids contact with Muslim groups. Muslims, in turn, experience severe disadvantages in the labour market, and even lag behind other minority groups in access to employment, wages, and occupational status (Sobolewska, Galandini, & Lessard-Phillips, 2017; Weichselbaumer, 2017). Muslim migrants in Western Europe tend to originate from a small cluster of countries (e.g., Morocco, Turkey, Pakistan, Somalia), which raises the question whether the disadvantage they experience is due to their national origin or their religion. In the GEMM study, the orthogonal variation of ethnicity and religion enables us to tackle this identification problem (Di Stasio et al., 2019).

Fig. 4 plots the effect of signaling religion in the application for migrants who only differ in religious affiliation. For this analysis, we only focused on origin countries where both Muslims and Christians are sizeable groups (Albania, Bulgaria, Egypt, Ethiopia, Indonesia, Lebanon, Nigeria, Russia, Uganda). Clearly, it is not mentioning religion *per se* that puts applicants at a disadvantage. Employers are only penalizing applicants with a Muslim background. Applicants of the *same ethnic origin* but signaling closeness to Christianity rather than Islam are treated just as well as applicants who do not disclose their religion. It is only with a factorial design independently varying country of origin and religion that Islamophobia can be singled out as one of the drivers of discrimination.

5. Discussion and conclusion

In this paper, we have identified three gaps in the existing (field experimental) literature on ethnic discrimination and proposed three

[7] Note that for this analysis some minority groups (e.g. Indians, Latin Americans, Caribbeans) were dropped from the sample.

methodological innovations in the design of field experiments that offer analytical leverage to better understand the scope conditions, drivers and targets of discrimination. We have relied on data from the GEMM study, a large-scale field experiment conducted in five European countries, to show the potential of using cross-national, multi-group and factorial designs to answer important questions on where, why and against whom employers discriminate.

More specifically, we have stressed how a cross-national design with direct equivalence in ethnic groups and occupations may be necessary to identify the contexts where minorities suffer the greatest disadvantage. We have argued that multi-group designs can offer new insight into whether employers' reluctance to hire minority groups can be explained by ethnic homophily (i.e. a general preference for the ingroup), or are patterned along a gradual ethnic hierarchy. The inclusion of many origin groups varying in cultural distance, race, religion, colonial ties with the host country, etc., offers analytical leverage to explain how these hierarchies are formed. Finally, to understand the mechanisms behind employers' preferences, we believe factorial designs are a promising strategy. To give an example, we have shown how the orthogonal variation of religion and country of origin provides compelling evidence that employers are, at least partly, driven by Islamophobia.

We have paid particular attention to the issue of external validity. While we agree with Lahey and Beasley (Lahey & Beasley, 2018, p. 82) that "ultimately, the external validity of an experiment is constrained by each decision made in the design", we have elaborated on those issues that were of particular concern to us while working on the GEMM study: the choice of which origin groups and destination countries to study, the characteristics that are randomly varied in the experimental design in addition to ethnicity, the different strength of the ethnicity signal in labor markets that have very different application standards, and how to choose occupations that can guarantee a consistent volume of job openings for the entire duration of fieldwork.

While we believe that the GEMM study is an important contribution to the field of comparative experimental research on discrimination, naturally, the study also has its limitations, which could be addressed in future research projects. We focus here on limitations that are particularly relevant for cross-national comparisons. First, only one name per ethnic minority group was used. However, the level of discrimination recorded in field experiments may be affected by the specific names used to signal ethnicity or race. Names intended to unequivocally signal ethnic background may convey information on other variables such as socio-economic status (Dahl & Krog, 2018). Moreover, the perception or recognition of names may differ across countries; as an example, recent field experiments found that Nigerian names in Austria (Weichselbaumer, 2017) and Caribbean names in Britain (Wood et al., 2009) were especially hard to recognize (see also Gaddis, 2017 for a discussion on names).

A second limitation has to do with the use of different online job boards to sample job openings in the five countries. These websites varied in the amount and quality of information provided on the jobs being advertised. As a result, while we did our best to harmonize the job applications across countries, the available information on employers and on the sampled jobs is country-specific and, therefore, poorly comparable. Our strategy was to keep track of all information available in each national context so that the organizational drivers of discrimination could still be analyzed in single-country studies.

A more general limitation of field experiments is that there usually is only 'one' job applicant profile: in the GEMM study, the applicant is an individual aged 22–26 years old, who is fulltime employed with four years of working experience. Yet, discrimination is likely to be different for older people, or people with a different employment history (Pedulla, 2016; Riach & Rich, 2010). More importantly, these differences might vary across countries. For example, the rank order of the discrimination ratios across countries might differ by applicant's age. To some extent, our choice to use a factorial design and vary a number of other applicants' characteristics besides ethnicity minimizes this template bias.

The methodological innovations that we introduced in the GEMM study are by no means unique to field experiments. Double comparative designs can be applied in research projects that use other experimental methods, such as, for example, survey experiments. Cross-national survey experiments on immigrants' integration are still rare; existing works examine different origin groups in different countries (Kootstra, 2016; Sobolewska et al., 2017; Valentino et al., 2017), favoring functional over direct equivalence. Survey and field experiments with complex factorial designs might even be combined in a multi-method study to better understand the beliefs and stereotypes underlying employers' biases (Gaddis, 2019).

We do not consider one method as intrinsically better than the other. While cross-national field experiments are desirable for their strong internal and external validity, they are also time- and labor-intensive to conduct, especially in countries like Germany or Switzerland where application standards are very strict and copies of certificates or reference letters are required. Design harmonization across countries that differ substantially in application standards can be problematic. Realistically, cross-national field experimental designs can only include a handful of countries, which reduces the analytical leverage to test institutional explanations. For these reasons, cross-national designs should be based on a careful case selection, possibly opting for designs that offer analytical leverage for theory testing, such as most-likely or least-likely designs (Levy, 2008, p. 12).

If cross-national comparisons are not an option, even in single-country field experimental studies, a preference for multi-group and factorial designs would yield estimates of discrimination that are more externally valid. These estimates can then be analyzed and compared using meta-analysis, which remains a powerful tool to put more focused comparisons based on cross-national field experiments into perspective. Considerable resources are needed to carry out a double-comparative, factorial field experiment. Yet, we see many analytical advantages of using such data; we therefore encourage future researchers to implement some of the three methodological innovations in their research designs.

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.2019.100463.

References

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*(3), 237–269. https://doi.org/10.1177/0730888412444783.

Baldassarri, D., & Abascal, M. (2017). Field experiments across the social sciences. *Annual Review of Sociology, 43*(1), 41–73. https://doi.org/10.1146/annurev-soc-073014-112445.

Becker, S. O., Fernandes, A., & Weichselbaumer, D. (2019). Discrimination in hiring based on potential and realized fertility: Evidence from a large-scale field experiment. *Labour Economics, 59*, 139–152. https://doi.org/10.1016/j.labeco.2019.04.009.

Bertrand, M., & Duflo, E. (2017). Chapter 8—Field experiments on discrimination. In A. V. Banerjee, & E. Duflo (Eds.). *Handbook of economic Field experiments* (pp. 309–393). . https://doi.org/10.1016/bs.hefe.2016.08.004.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*(4), 991–1013.

Bessudnov, A., & Shcherbak, A. (2019). Ethnic discrimination in multi-ethnic societies: Evidence from Russia. *European Sociological Review*. https://doi.org/10.1093/esr/

jcz045.
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*(4), 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.
Brinton, M. C., & Nee, V. (1998). *The new institutionalism in sociology.* New York: Russell Sage Foundation.
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*(3), 399–409. https://doi.org/10.1093/esr/jcu047.
Carbonaro, W., & Schwarz, J. (2018). Opportunities and challenges in designing and conducting a labor Market resume study. In S. M. Gaddis (Ed.). *Audit studies: Behind the scenes with theory, method, and nuance* (pp. 143–158). . https://doi.org/10.1007/978-3-319-71153-9_7.
Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review, 43*(6), 1241–1299. https://doi.org/10.2307/1229039.
Crul, M. (2016). Super-diversity vs. assimilation: How complex diversity in majority–Minority cities challenges the assumptions of assimilation. *Journal of Ethnic and Migration Studies, 42*(1), 54–68. https://doi.org/10.1080/1369183X.2015.1061425.
Dahl, M., & Krog, N. (2018). Experimental evidence of discrimination in the labour market: Intersections between ethnicity, gender, and socio-economic status. *European Sociological Review, 34*(4), 402–417. https://doi.org/10.1093/esr/jcy020.
Di Stasio, V., Lancee, B., Veit, S., & Yemane, R. (2019). Muslim by default or religious discrimination? Results from a cross-national field experiment on hiring discrimination. *Journal of Ethnic and Migration Studies*.
Eden, D. (2017). Field experiments in organizations. *Annual Review of Organizational Psychology and Organizational Behavior, 4*(1), 91–122. https://doi.org/10.1146/annurev-orgpsych-041015-062400.
Edo, A., Jacquemet, N., & Yannelis, C. (2019). Language skills and homophilous hiring discrimination: Evidence from gender and racially differentiated applications. *Review of Economics of the Household, 17*(1), 349–376. https://doi.org/10.1007/s11150-017-9391-z.
Gaddis, S. M. (2017). Racial/Ethnic perceptions from hispanic names: Selecting names to test for discrimination. *Socius, 3*. https://doi.org/10.1177/2378023117737193 2378023117737193.
Gaddis, S. M. (2019). Understanding the "How" and "Why" aspects of racial-ethnic discrimination: A multimethod approach to audit studies. *Sociology of Race and Ethnicity*. https://doi.org/10.1177/2332649219870183 2332649219870183.
Goldthorpe, J. H. (2016). *Sociology as a population science.* Cambridge University Press.
Hagendoorn, L. (1995). Intergroup biases in multiple group systems: The perception of ethnic hierarchies. *European Review of Social Psychology, 6*(1), 199–228. https://doi.org/10.1080/14792779443000058.
Heath, A., & Di Stasio, V. (2019). Racial Discrimination in Britain, 1969-2017: A meta-analysis of field experiments on racial discrimination in the labour market. Forthcoming *The British Journal of Sociology*.
Heath, A. (2013). Can religious affiliation explain "ethnic"inequalities in the labour market? *Ethnic and Racial Studies, 36*(6), 1005–1027.
Heath, A., Rothon, C., & Kilpi, E. (2008). The second generation in western Europe: Education, unemployment, and occupational attainment. *Annual Review of Sociology, 34*, 211–235.
Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods, 12*(3), 554–566. https://doi.org/10.1177/1094428107300396.
Jackson, M., & Cox, D. R. (2013). The principles of experimental design and their application in sociology. *Annual Review of Sociology, 39*(1), 27–49. https://doi.org/10.1146/annurev-soc-071811-145443.
Jacquemet, N., & Yannelis, C. (2012). Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market. *Labour Economics, 19*(6), 824–832. https://doi.org/10.1016/j.labeco.2012.08.004.
Johnson, T. P. (1998). *Approaches to equivalence in cross-cultural and cross-national survey research.* DEU.
Kaas, L., & Manger, C. (2012). Ethnic discrimination in Germany's labour market: A field experiment. *German Economic Review, 13*(1), 1–20. https://doi.org/10.1111/j.1468-0475.2011.00538.x.
King, E. B., Hebl, M. R., Botsford Morgan, W., & Ahmad, A. S. (2013). Field experiments on sensitive organizational topics. *Organizational Research Methods, 16*(4), 501–521. https://doi.org/10.1177/1094428112462608.
Kogan, I. (2006). Labor markets and economic incorporation among recent immigrants in Europe. *Social Forces, 85*(2), 697–721.
Koopmans, R., Veit, S., & Yemane, R. (2018). *Ethnische Hierarchien in der Bewerberauswahl: Ein Feldexperiment zu den Ursachen von Arbeitsmarktdiskriminierung.* https://doi.org/SP VI 2018-104.
Koopmans, R., Veit, S., & Yemane, R. (2019). Taste or statistics? A correspondence study of ethnic, racial and religious labour market discrimination in Germany. *Ethnic and Racial Studies, 42*(16), 233–252. https://doi.org/10.1080/01419870.2019.1654114.
Kootstra, A. (2016). Deserving and undeserving welfare claimants in Britain and the Netherlands: Examining the role of ethnicity and migration status using a vignette experiment. *European Sociological Review, 32*(3), 325–338. https://doi.org/10.1093/esr/jcw010.
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). . https://doi.org/10.1007/978-3-319-71153-9_4.
Lahey, J., & Beasley, R. A. (2009). Computerizing audit studies. *Journal of Economic Behavior & Organization, 70*(3), 508–514. https://doi.org/10.1016/j.jebo.2008.02.009.
Lancee, B. (2016). The negative side effects of vocational education: A cross-national analysis of the relative unemployment risk of young non-western immigrants in Europe. *The American Behavioral Scientist, 60*(5–6), 659–679. https://doi.org/10.1177/0002764216632835.
Lancee, B. (2019). Ethnic discrimination in hiring: Comparing groups across contexts. Results from a cross-national field experiment. *Journal of Ethnic and Migration Studies*. https://doi.org/10.1080/1369183X.2019.1622744.
Lancee, B., Birkelund, G., Coenders, M., Di Stasio, V., Fernandez Reino, M., Heath, A., ... Soinee, H. (2019). *The GEMM study: A cross-national harmonized field experiment on labour market discrimination-Codebook.* https://doi.org/10.2139/ssrn.3398273 Available at SSRN 3398273.
Lancee, B., Birkelund, G., Coenders, M., Di Stasio, V., Fernandez Reino, M., Heath, A., ... Thijssen, L. (2019). *The GEMM study: A cross-national harmonized field experiment on labour market discrimination: Technical report*https://doi.org/10.2139/ssrn.3398191 Available at SSRN 3398191.
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.
Levy, J. S. (2008). *Case studies: Types, designs, and logics of inference: Conflict management and peace science.* https://doi.org/10.1080/07388940701860318.
Mccall, L. (2008). *The complexity of intersectionality.* August 21https://doi.org/10.4324/9780203890882-11.
Midtbøen, A. (2015). The context of employment discrimination: Interpreting the findings of a field experiment. *The British Journal of Sociology, 66*(1), 193–214. https://doi.org/10.1111/1468-4446.12098.
Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature, 56*(3), 799–866. https://doi.org/10.1257/jel.20161309.
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*(4), 148–171. https://doi.org/10.1257/pol.3.4.148.
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.
Pager, D., Western, B., & Bonikowski, B. (2009). Discrimination in a low-wage labor market. *American Sociological Review, 74*(5), 777–799. https://doi.org/10.1177/000312240907400505.
Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology, 34*, 181–209.
Pedulla, D. S. (2016). Penalized or protected? Gender and the consequences of non-standard and mismatched employment histories. *American Sociological Review, 81*(2), 262–289.
Pedulla, D. S. (2018). Emerging frontiers in audit study research: Mechanisms, variation, and representativeness. In S. M. Gaddis (Ed.). *Audit studies: Behind the scenes with theory, method, and nuance* (pp. 179–195). . https://doi.org/10.1007/978-3-319-71153-9_9.
Polavieja, J. G., Lancee, B., Ramos, M., Veit, S., & Yemane, R. (2019). *Phenotypic discrimination in hiring: Results from a comparative field experiment in Germany, the Netherlands and Spain.* Working PaperGEMM.
Quillian, L., Heath, A., Pager, D., Midtbøen, A. H., Fleischmann, F., & Hexel, O. (2019). Do some countries discriminate more than others? Evidence from 97 field experiments of racial discrimination in hiring. *Sociological Science, 6*, 467–496.
Riach, P. A. (2015). A field experiment investigating age discrimination in four European labour markets. *International Review of Applied Economics, 29*(5), 608–619. https://doi.org/10.1080/02692171.2015.1021667.
Riach, P. A., & Rich, J. (2010). An experimental investigation of age discrimination in the english labor market. *Annals of Economics and Statistics,* (99/100), 169–185. https://doi.org/10.2307/41219164.
Ross, S. (2017). Measuring trends in discrimination with field experiment data. *Proceedings of the National Academy of Sciences, 114*(41), 10815–10817.
Sobolewska, M., Galandini, S., & Lessard-Phillips, L. (2017). The public view of immigrant integration: Multidimensional and consensual. Evidence from survey experiments in the UK and the Netherlands. *Journal of Ethnic and Migration Studies, 43*(1), 58–79. https://doi.org/10.1080/1369183X.2016.1248377.
Thijssen, L., Lancee, B., Veit, S., & Yemane, R. (2019). Discrimination against Turkish minorities in Germany and the Netherlands: Field experimental evidence on the effect of diagnostic information on labour market outcomes. *Journal of Ethnic and Migration Studies*. https://doi.org/10.1080/1369183X.2019.1622793.
Valentino, N. A., Soroka, S. N., Iyengar, S., Aalberg, T., Duch, R., Fraile, M., et al. (2017). Economic and cultural drivers of immigrant support worldwide. *British Journal of Political Science,* 1–26. https://doi.org/10.1017/S000712341700031X.
Van De Vijver, F. J., & Poortinga, Y. H. (1982). Cross-cultural generalization and universality. *Journal of Cross-cultural Psychology, 13*(4), 387–408.
Van Tubergen, F., Maas, I., & Flap, H. D. (2004). The economic incorporation of immigrants in 18 western societies: Origin, destination and community effects. *American Sociological Review, 69*(5), 704–727.
Vertovec, S. (2007). Super-diversity and its implications. *Ethnic and Racial Studies, 30*(6),

1024–1054. https://doi.org/10.1080/01419870701599465.
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). . https://doi.org/10.1007/978-3-319-71153-9_6.
Weichselbaumer, D. (2017). Discrimination against migrant job applicants in Austria: An experimental study. *German Economic Review, 18*(2), 237–265.
Wood, M., Hales, J., Purdon, S., Sejersen, T., & Hayllar, O. (2009). *A test for racial discrimination in recruitment practice in British cities*Norwich: Department for Work and Pensions Research Report 607.
Zegers de Beijl, R. (Ed.). (2000). *Documenting discrimination against migrant workers in the labour market. A comparative study of four European countries*. Geneva: ILO.
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.