European Economic Review 151 (2023) 104315

Contents lists available at ScienceDirect

European Economic Review

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

The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments

Louis Lippens [a,b,*], Siel Vermeiren [a], Stijn Baert [a,c]

[a] Ghent University
[b] Vrije Universiteit Brussel
[c] University of Antwerp; Université catholique de Louvain; Institute of Labor Economics (IZA); Global Labor Organization (GLO)

ARTICLE INFO

JEL classification:
J71
J23
J14
J15
J16

Keywords:
Hiring discrimination
Unequal treatment
Meta-analysis
Correspondence experiment
Audit study

ABSTRACT

Notwithstanding the improved integration of various minority groups in the workforce, unequal treatment in hiring still hinders many individuals' access to the labour market. To tackle this inaccessibility, it is essential to know which and to what extent minority groups face hiring discrimination. This meta-analysis synthesises a quasi-exhaustive register of correspondence experiments on hiring discrimination published between 2005 and 2020. Using a random-effects model, we computed pooled discrimination ratios concerning ten discrimination grounds upon which unequal treatment in hiring is forbidden by law. Our meta-analysis shows that hiring discrimination against candidates with disabilities, older candidates, and less physically attractive candidates seems equally severe as the unequal treatment of candidates with salient racial or ethnic characteristics. Moreover, hiring discrimination against older applicants is more prominent in Europe than in the United States. Last, while we initially find a significant decrease in ethnic hiring discrimination in (Western) Europe, we find no structural evidence of recent temporal changes in hiring discrimination when controlling for the minority groups considered, at the country level, or based on the various other grounds within the scope of this review.

1. Introduction

Although the workforce has become increasingly diverse—improving the integration of female, migrant, and older workers, amongst other groups—many individuals belonging to various minority groups still face considerable discrimination in the labour market (Organisation for Economic Co-operation and Development [OECD], 2020a). In part because of their decreased chances for labour market access, these individuals are at elevated risk of long-term unemployment and labour market inactivity (OECD, 2020a, 2020b). This underutilisation of talent could result in needless economic costs for firms and society (Baert, 2021; OECD, 2020a; Pager, 2016). For policymakers, it is vital to know which (minority) groups are confronted with hiring discrimination and to understand the severity of this labour market's inaccessibility. In this way, targeted diversity policies, such as outreach campaigns focusing on underrepresented or discriminated groups, can be implemented to help those who require said policies the most (OECD, 2020a).

Research on labour market discrimination has long focused on the non-experimental decomposition approach to measure discrimination (Blinder, 1973; Kitagawa, 1955; Neumark, 2018; Oaxaca, 1973). This approach has historically involved isolating the

* Corresponding author at: Louis Lippens, Faculty of Economics and Business Administration, Ghent University, Sint-Pietersplein 6, 9000 Ghent, Belgium; louis.lippens@ugent.be
E-mail address: louis.lippens@ugent.be (L. Lippens).

https://doi.org/10.1016/j.euroecorev.2022.104315
Received 17 January 2022; Received in revised form 31 August 2022; Accepted 2 October 2022
Available online 20 October 2022
0014-2921/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

impact of discrimination on wages via regression analyses (Borjas, 2020). Variance that could not be explained by differences in human capital between groups of interest (e.g. Blacks and Whites) was consequently attributed to discrimination. However, it is difficult to capture the true amount of variance explained by human capital under this approach, primarily due to omitted variable bias (Altonji and Blank, 1999; Borjas, 2020).[1] The decomposition method thus sketches an incomplete picture of discrimination (Borjas, 2020; Gaddis, 2018).

To overcome the limitation of the decomposition approach, researchers began to use audit studies as an alternative experimental method to measure the incidence of labour market discrimination (Gaddis, 2018). At first, this was mainly done by sending out pairs of real applicants (i.e. actors) who differed in terms of visible characteristics based on which unequal treatment is forbidden (e.g. skin colour) to interview for the same job. Differences in job offers were subsequently interpreted in terms of discrimination. In the early 2000s, however, Bertrand and Mullainathan (2004) steered the research area of labour market discrimination in a different direction: correspondence audits replaced in-person audits as the standard for measuring hiring discrimination (Gaddis, 2018).[2] Rather than sending out actors as applicants, these correspondence experiments consisted of mailing written applications from fictitious job seekers in response to real job postings. By randomly assigning individual characteristics based on which selection is forbidden, the effect of these characteristics on employers' reactions can be given causal interpretations. Compared with in-person audits, the perceived differences between applicants, produced by minute differences in their behaviour during the interview, are nullified. Moreover, the application process is less resource-intensive. Because of its high employability and the causal interpretation that underpins its results, the correspondence testing method is still considered the reference method for measuring hiring discrimination at present (Baert, 2018; Neumark, 2018; Verhaeghe, 2022).

In recent years, a considerable number of scholars have reviewed and synthesised parts of the hiring discrimination literature, focusing on the correspondence testing method. We know of nineteen topical meta-studies in this evolving research area: Adamovic (2020, 2022), Baert (2018); Bartkoski et al. (2018); Batinović et al. (2022); Bertrand and Duflo (2017); Derous and Ryan (2019), Flage, (2020), Gaddis et al. (2021), Heath and Di Stasio (2019), Lippens et al. (2022), Neumark (2018), Quillian and Midtbøen (2021), Quillian et al. (2017, 2019, 2020), Rich (2014), Thijssen et al. (2021), and Zschirnt and Ruedin (2016). Table A1 in Appendix A provides an overview of the aforementioned meta-studies including details about the (i) type of review, (ii) type of analysis, (iii) inclusion of a meta-regression component, (iv) included discrimination grounds, (v) type of studies considered, (vi) region, (vii) period, and (viii) main findings.[3] These studies can be roughly classified into two categories: traditional reviews and systematic reviews (Briner and Denyer, 2012). Traditional reviews, on the one hand, typically lack structure in the scoping or search process; it is not always clear why some studies are highlighted while others are ignored. Few recent meta-studies can be categorised as such. Systematic reviews, on the other hand, have a clear scope and target specific studies meeting predetermined inclusion criteria. The majority of recent meta-studies summarising correspondence audits are systematic reviews of which most also pursue meta-analytic methods to quantify and contextualise hiring discrimination.

All of these meta-studies have merit. Several of these studies have provided insightful empirical, theoretical or policy-orientated observations but, for understandable reasons, have only focused on specific grounds of discrimination (predominantly race, ethnicity, and national origin) while neglecting others (e.g. Bartkoski et al., 2018; Gaddis et al., 2021; Quillian et al., 2017). Other reviews taking a broader view of hiring discrimination have brought forth equally interesting insights but do not provide a systematic account of the existing correspondence experiment literature (e.g. Bertrand and Duflo, 2017; Neumark, 2018; Rich, 2014). Moreover, recent meta-studies have successfully implemented meta-regression techniques to identify interesting patterns in the data. For example, Gaddis et al. (2021) found that controlling for relevant covariates, discrimination against Black Americans is highest in high-stake settings such as hiring and housing (versus education, medical services, or public services). Quillian et al. (2019) found that there is less discrimination in jobs where a college degree is required (versus a high school degree or equivalent). Yet, again, the scope of these studies has been limited to race, ethnicity, and national origin. Baert (2018) was the first in attempting to counter these limitations by (i) adopting a broad view on hiring discrimination considering all grounds based on which unequal treatment is forbidden under United States federal and state law and (ii) providing a quasi-exhaustive register of correspondence experiments conducted since Bertrand and Mullainathan's (2004) seminal study. However, the main limitation of Baert's (2018) work is the absence of meta-analysis to synthesise and compare the results of the included studies.

The current study documents the most extensive register of correspondence experiments on hiring discrimination to date. We compile and synthesise a comprehensive catalogue of correspondence experiments published between 2005 and 2020, predominantly from the Americas, Europe, and Asia. We simultaneously provide an overview of the presently under-researched grounds for discrimination and the hiatuses that still exist in the current literature on hiring discrimination. Altogether, we gather 306 correspondence experiments (i.e. units of observation in our analysis) originating from 169 separate correspondence audit studies. Adding up the job applications from each correspondence experiment, these experiments comprise almost 965,000 fictitious applications in response to real job vacancies.

[1] Classic examples of omitted variables include (but are not limited to) unobserved supply-side factors such as personal motivation or choices as well as possessing an extensive professional network.

[2] Bertrand and Mullainathan (2004) were not the first to implement the correspondence audit method (see e.g. Jowell & Prescott-Clarke, 1970). However, their influential study gave traction to the so-called 'third wave' (and the current fourth wave) of audit studies which mainly comprised correspondence audits (Gaddis, 2018). Since then, the publication density of correspondence audits has increased substantially.

[3] Other meta-studies that do not report on audits studies related to hiring discrimination were not considered for this overview. One example thereof is Drydakis (2022), who recently examined the literature on the relationship between sexual orientation and earnings differences.

Our synthesis offers scholars and policymakers an understanding of the prevalence and severity of hiring discrimination concerning various discrimination grounds within the scope of this review. Specifically, we meta-analytically quantify hiring discrimination regarding ten discrimination grounds upon which unequal treatment is forbidden under United States federal or state law, including (i) race, ethnicity, and national origin, (ii) gender and motherhood status, (iii) age, (iv) religion, (v) disability, (vi) sexual orientation, (vii) physical appearance, (viii) wealth, (ix) marital status, and (x) military service or affiliation. The standardised meta-analytical approach enables comparisons of levels of hiring discrimination across discrimination grounds and minority groups.[4] We also assess heterogeneity in hiring discrimination, providing a more granular perspective of our findings. For each discrimination ground, we explore whether (i) levels of hiring discrimination are related to how call-backs are reported and measured, (ii) persistent regional differences in hiring discrimination exist, or (iii) unequal treatment in hiring has changed in recent times. In summary, this is the first study using meta-analytic techniques to make such a broad comparison of previous experiments on hiring discrimination, across discrimination grounds and minority groups, based on the currently largest documented set of correspondence audit studies from across the globe.

2. Data and methods

In this section, we elaborate on (i) the scope of our meta-analysis; (ii) how we identified and selected studies; (iii) which variables we collected from these studies, as well as how we classified some of them into broader categories to identify differences across these categories; and (iv) the details of the meta-analytical methods we used to analyse the resulting data. In this process, we paid special attention to the reporting guidelines for meta-analyses in economics of Havránek et al. (2020)—we refer the reader to Appendix A (Table A2) for the corresponding checklist.

2.1. Scope

We used various eligibility criteria based on the Population, Intervention, Comparison, Outcome (PICO) framework to delineate our review (Richardson et al., 1995).[5] Table 1 provides an overview of these criteria. We limited our review to correspondence studies in which unequal treatment in hiring was assessed between fictitious applicants belonging to minority groups and their majority counterparts. We considered studies written in English that were first published as a discussion paper, pre-print, or journal article between 2005 (the year after Bertrand and Mullainathan's seminal 2004 correspondence study) and 2020 (the most recent full calendar year at the time this study was conducted) in particular.[6]

Similar to the delineation of the discrimination grounds in Baert's (2018) correspondence experiment register, we limited our scope to the forms of hiring discrimination prohibited under United States federal or state law.[7] We focused on United States federal and state law for two reasons. First, we wanted to ensure the complementarity of our dataset with Baert (2018). Their original reasoning was to consider legal discrimination grounds where most correspondence audit studies (focusing on hiring discrimination) were conducted—this was (and still is) the United States. Second, compared to the European Union, where many audit studies originate from, there is less room for the discretionary application of employment discrimination law compared to the United States (Ganty and Benito Sanchez, 2021). We thus took into account the following discrimination grounds: (i) race, ethnicity, and national origin, (ii) gender and motherhood status, (iii) religion, (iv) disability, (v) age, (vi) military service or affiliation, (vii) wealth, (viii) genetic information, (ix) citizenship status, (x) marital status, (xi) sexual orientation, (xii) political affiliation, (xiii) union affiliation, and (xiv) physical appearance. In our final analysis, we retained only ten of these grounds because, similar to Baert (2018), (i) no (new) correspondence experiments related to genetic information or citizenship status were identified in the search process; and (ii) we found only one experiment related to political orientation and one experiment related to union affiliation, from which we could not calculate pooled discrimination ratios.

2.2. Study selection

We used multiple sources to identify, screen, and select eligible studies for our meta-analysis. Fig. 1 depicts a structured overview of this process. First, we identified potentially eligible correspondence studies. On the one hand, we sourced studies included in Baert's (2018) register of correspondence experiments, which resulted from an elaborate systematic search for correspondence experiments on hiring discrimination. On the other hand, we performed a systematic search on the Web of Science and Google Scholar databases in spring 2021. Our search used the keywords 'correspondence experiment', 'correspondence study', 'fictitious resume', 'fictitious cv',

[4] We do not attempt to provide a reasoning for the underlying mechanisms of the uncovered hiring discrimination. Our primary goal remained to compare levels of hiring discrimination across discrimination grounds and minority groups. For recent overview studies on the empirical evidence concerning the economic mechanisms of (ethnic) labour market discrimination, we refer the reader to Lang & Kahn-Lang Spitzer (2020), Lippens et al. (2022), and Neumark (2018).

[5] We extended the PICO framework to be more specific in the delineation of the scope of our review. Precisely, we also considered 'study type', 'context', and 'timing' and excluded 'intervention' because it was not relevant to our search query.

[6] The specific year allocated to a given study was based on the year the study was initially published. For example, Larsen and Di Stasio (2021) was first published online as a pre-print in 2019 before appearing officially in the *Journal of Ethnic and Migration Studies* in 2021.

[7] A reassessment of United States federal and state law was made in late 2020. Relative to Baert (2018), not much had changed. One noteworthy change, however, is that discrimination based on LGBT+ status has been made illegal at the US federal level (Boystock v. Clayton County, 2020).

Table 1
Eligibility criteria for study inclusion.

Criterion	Details
Study type	Correspondence experiment in which applications were sent in response to vacancies.
Population	(Fictitious) applicants from various minority groups and their majority counterparts.
Outcome	Disadvantageous, unequal treatment in the hiring and selection process (i.e. hiring discrimination).
Comparison	Hiring chances of minority applicants compared with those of majority applicants.
Context	Hiring discrimination related to the grounds upon which unequal treatment is forbidden under United States federal or state law (i.e. race, ethnicity, and national origin, gender and motherhood status, religion, disability, age, military service or affiliation, wealth, genetic information, citizenship status, marital status, sexual orientation, political orientation, union affiliation, and physical appearance).
Timing	Study first published between 2005 and 2020.

Notes. The framework used to define the eligibility criteria is based on the PICO (Population, Intervention, Comparison, Outcome) framework first coined by Richardson et al. (1995).

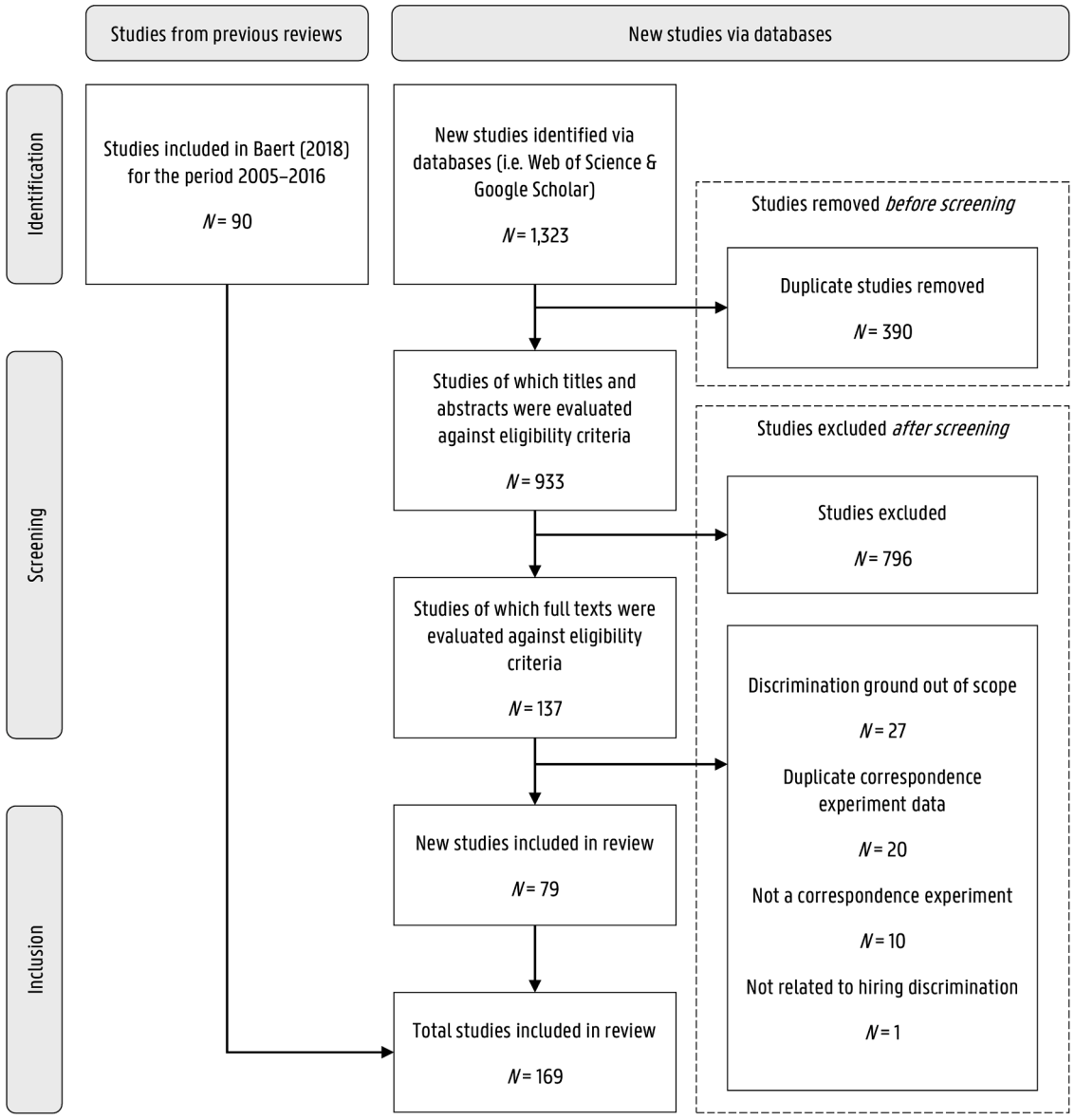

Fig. 1. Study selection flow diagram. Notes. This figure is adapted from Page et al., 5).

'fictitious application', and 'field experiment' in combination with the keyword 'discrimination'. In general, we confined our search to studies published in the period 2005 to 2020. To extend this systematic search, we also performed a cited reference search with the references from Baert's (2018) book chapter as the input of our queries

Next, we appraised the studies that had not already been identified by Baert (2018). In total, we evaluated the titles and abstracts of 933 studies against our eligibility criteria (see Section 2.1). After an initial screening of the titles and abstracts, we reviewed the full text of the remaining 137 articles. The risk of reviewer bias was reduced by having two researchers independently review the selected articles. After this review, 79 studies were identified that fully matched the criteria. There were four reasons for excluding certain studies after appraising their full text: (i) unequal treatment based on the discrimination ground in the scope of the study was not forbidden under United States law ($N = 27$, 46.55% of the total number of excluded full texts);[8] (ii) the correspondence experiment was entirely based on data used in a previously published (and already included) study ($N = 20$, 34.48%);[9] (iii) the study did not use the correspondence testing method (e.g. in-person audit; $N = 10$, 17.24%); or (iv) the study was solely related to housing discrimination instead of hiring discrimination ($N = 1$, 1.72%).

We retained a total of 169 studies, of which 90 were already included in Baert's (2018) book chapter, resulting in 306 units of observation. There are more units of observation than studies due to our definition of a 'unit of observation', i.e. a unique correspondence experiment based on the related (i) discrimination ground, (ii) treatment group, (iii) control group, and (iv) region where the test was performed. For example, Di Stasio et al. (2021) considered hiring discrimination against Muslims in Germany, Norway, Spain, the Netherlands, and the United Kingdom. To allow for heterogeneity analyses on the basis of region (see Sections 2.4.2 and 3.4.2), this study was subdivided into multiple units of observation stemming from the same study.

2.3. Data collection

We captured a multitude of variables for each correspondence experiment. First, we registered the basic information of the studies, including the authors' names and the year the article was officially published. In addition to the latter, we also recorded (i) the year the study was initially published (e.g. as a pre-print or early-access article), which was the year we used when evaluating the article against our eligibility criteria, and (ii) the year the correspondence experiment ended.[10] Second, we documented where the research took place, including the country and (sub-)region. The latter was based on the M49 Standard for geographic regions of the United Nations (2021; see Table A3 for a tabulated overview). Third, we registered the (experimental) treatment group and the control group of the correspondence experiment. The specific treatment groups identified in the included studies were classified into broader groups to facilitate further analyses. Because no common global framework of ethnic and racial minority groups exists, the classification of these groups consisted of a proprietary framework based on how various governmental bodies of OECD member countries collect and categorise diversity data (Balestra and Fleischer, 2018; European Commission, 2021; Morning, 2008).[11] The classification related to the other discrimination grounds was based on the logical grouping of the respective treatment groups. The final classification can be found in Appendix A (Table A4).

Fourth, we documented data related to the outcomes of the correspondence experiments. We captured the overall treatment effect of the results in the original studies (averaged across sub-groups at the experiment level). We also recorded the classification of the outcome variable (i.e. call-back). If a call-back consisted of an invitation to a job interview (or any broadly defined positive response of the employer, such as a request for additional information), we labelled it narrow (or broad).

Most importantly, we registered the number of observations (i.e. fictitious job applications) and the number of positive call-backs in both the treatment and control groups. The accuracy of these variables was independently assessed and verified by at least two authors. Outcome measure data required to calculate pooled discrimination ratios were missing for 32 experiments (9.82%, out of a total of 326 units). After contacting the corresponding authors of the respective studies to retrieve these data, 12 cases could be completed (37.50% of cases with missing data), meaning that we had no data for the remaining 20 units of observation.[12] These units were excluded from the meta-analysis, resulting in 306 valid units (cfr. supra). Reporting bias, which could (partly) originate from missing data, was formally evaluated when we tested for publication bias (see Section 2.4.3).

From these data, we derived a standardised discrimination ratio. The specification of this ratio is shown in Eq. (1). The

[8] For example, Gaddis, (2015) looks at unequal treatment based on the educational institution an applicant attended. Making selection decisions based on the educational institution is not a ground for (illegal) discrimination. While the author makes an interesting assessment by looking at the interaction of this criterion with race, the results from this particular part of the study were not included in our analyses.

[9] Some studies we initially identified were based on data already contained in other studies. Including identical data multiple times in our analyses would obviously bias the pooled discrimination ratios. Simultaneously, this would lower the variance around these estimates. To avoid this 'multiple publication bias', we excluded such studies from our review (see Page et al., 2020).

[10] There is a discrepancy between the temporal period of this review (2005–2020) and the timeframe used in the heterogeneity analyses (see section 3.1 and 3.4.2). This is because the latter is based on the year in which the correspondence experiment ended. The rationale for this is that this time variable more accurately represents the timing of the experiment (vis-à-vis the year the research was published).

[11] In their correspondence test, Jacquement and Yannelis (2012), for example, assigned African American names to the minority group, while Gaddis, (2015) used Black-sounding names. In the United States and the United Kingdom, these origins are both classified as 'African (American)' or 'Black'. Therefore, we created the category 'African/African American/Black' as an umbrella term for similar treatment groups.

[12] These missing data were linked to ten studies: Beam et al. (2020), Carlsson and Eriksson (2019), Darolia et al. (2016), Drydakis (2017), Guul et al. (2019),Patacchini et al. (2015), Stone and Wright (2013), Thijssen, Coenders, & Lancee (2021), Thomas (2018), and Yemane and Fernández-Reino (2021).

discrimination ratio is a risk ratio (or relative risk) equal to the division of two proportions: (i) the proportion of positive call-backs in the treatment group (a_k) relative to the total number of observations in that group ($n_{k\ treat}$), and (ii) the proportion of positive call-backs in the control group (c_k) relative to the total number of observations in that group ($n_{k\ control}$). Because the discrimination ratio can be interpreted in terms of relative change, a ratio of 0.75, for example, indicates a 25% reduction in positive call-backs of the (fictitious) applicants of the minority group vis-à-vis the applicants of the control group, aggregated at the level of the correspondence experiment. Since our estimation strategy assumed that the included discrimination ratios follow a normal-like distribution, we log-transformed these ratios before pooling them in our meta-analysis (see Section 2.4.1). This approach ensured that opposite, same-sized effects were equidistant. In addition, Eq. (2) illustrates the calculation of the standard error of these log-transformed discrimination ratios (also see Harrer et al., 2021).

$$DR_k = \frac{a_k / n_{k\ treat}}{c_k / n_{k\ control}} \tag{1}$$

$$SE_{\ln DR} = \sqrt{\frac{1}{a_k} + \frac{1}{c_k} - \frac{1}{n_{k\ treat}} - \frac{1}{n_{k\ control}}} \tag{2}$$

2.4. Analyses

Our synthesis was based on the results of the correspondence experiments identified and selected in the previous steps. Our goals were (i) to quantify and compare the level of hiring discrimination for each of the various discrimination grounds and treatment groups in the scope of our analysis and (ii) to identify heterogeneity in hiring discrimination based on (a) the definition of the call-back variable, (b) the region where the correspondence experiment took place, and (c) the period of the experiment. We used R (version 4.1.0) for our analyses and relied on the {meta} package for most of our calculations (e.g. estimating the pooled ratios or detecting reporting bias; Balduzzi et al., 2019). We also used the {dmetar} package to identify influential cases, the {metasens} package to perform 'limit' meta-analyses, and the {metafor} package to perform meta-regression analyses and to examine the statistical (in) dependence of the sampled discrimination ratios (Harrer et al., 2019; Schwarzer et al., 2020; Viechtbauer, 2010).

2.4.1. Pooled discrimination ratios

To quantify the level of hiring discrimination across the various discrimination grounds, we used a random-effects model to pool the discrimination ratios of the included studies by discrimination ground and treatment group. We opted for this model because it starts from the premise that the true level of discrimination varies across studies. We assumed that there was at least some variation in these levels caused by (subtle) differences in the (i) definition and conceptualisation of the treatment and control groups, (ii) measurement of the responses (or call-backs), and (iii) overall experimental design and process. We applied Knapp–Hartung adjustments when calculating the confidence intervals around the pooled discrimination ratio estimates (Knapp and Hartung, 2003). This method assumes a *t*-distribution of the pooled effect rather than a normal distribution, which reduces the chance of obtaining false-positive results (Langan et al., 2019). The Knapp–Hartung adjustments produce more conservative (i.e. wider) confidence interval estimates than when these adjustments would not be applied.

To calculate the weights of the studies (w) in the reported pooled discrimination ratios, we used the commonly reported Mantel–Haenszel method for binary outcome data—the formula is shown in Eq. (3) (for more details, see Borenstein et al., 2009; Mantel and Haenszel, 1959). This method takes into account the number of cases in the treatment and control groups wherein the call-back was positive (a and c, respectively), as well as the number of cases in the treatment and control groups wherein the call-back was negative or absent (b and d, respectively; Mantel and Haenszel, 1959). This approach inherently attaches more importance to studies with larger sample sizes or overall higher numbers of positive call-backs. To generate more balanced weights, the weights were adjusted for between-study variance (τ^2). This procedure decreases potential overemphasis (or underemphasis) on studies with a relatively large (or small) sample size (see also Section 2.4.2; Borenstein et al., 2009). Subsequently, these variance-adjusted weights (w^*) were plugged into the general specification of the random-effects model, as illustrated in Eq. (4). Here, $\widehat{DR}$ is the pooled discrimination ratio, DR_k represents the observed discrimination ratio of the individual correspondence experiments, $\widehat{\zeta}$ is the error related to the overarching distribution of true discrimination ratios, and $\widehat{\varepsilon}$ symbolises the sampling error (Borenstein et al., 2009; Harrer et al., 2021).

$$w_k = \frac{(a_k + b_k) * c_k}{a_k + b_k + c_k + d_k} \tag{3}$$

$$\widehat{DR} = \frac{\sum_{k=1}^{K} (DR_k + \widehat{\zeta}_k + \widehat{\varepsilon}_k) * w_k^*}{\sum_{k=1}^{K} w_k^*} \tag{4}$$

2.4.2. Heterogeneity analyses

To meaningfully interpret the pooled discrimination ratios by discrimination ground and treatment group and to identify differences in hiring discrimination levels, we quantified and examined variability in statistical and design-related heterogeneity. First, we assessed statistical heterogeneity by calculating a statistic that captured the variability in the true discrimination ratios underlying the data (Rücker et al., 2008). More specifically, we calculated I^2 estimates, which indicate the proportion of the total variability due to

between-study variability (i.e. a value between 0 and 1) in the true discrimination ratios not caused by sampling error (Harrer et al., 2021; Higgins and Thompson, 2002; Veroniki et al., 2015). The I^2 statistic compares the studies' discrimination ratios to the pooled ratio, weighted by the inverse of the variance of the respective studies, taking into account the total number of studies. Therefore, this statistic is insensitive to (substantial) changes in the number of studies included in the analysis (Cochran, 1954; Harrer et al., 2021; Hoaglin, 2016). A high I^2 value (closer to 1) warrants exploring the source of this heterogeneity, for example by investigating moderation effects, as well as checking and controlling for possible extreme cases (i.e. outliers) in the dataset.[13]

Second, we evaluated design-related heterogeneity (i.e. heterogeneity due to differing designs across studies) by performing meta-regression analyses using the weighted least squares method (WLS) with a maximum-likelihood estimator. More specifically, we examined the heterogeneity in hiring discrimination for the following study-level variables: (i) the call-back classification, (ii) the geographical area where the correspondence experiment took place, and (iii) the year when the experiment ended. We did so largely in an exploratory manner—we tested how hiring discrimination contextually varied and compared our results with those of previous meta-studies where relevant. This approach contributed to (partly) explaining the statistical heterogeneity estimated in the previous step. Following Schwarzer et al., (2015) guidelines, we only performed meta-regressions on the groups of studies for which the total number of included studies was equal to or greater than ten.

The general meta-regression specification is given in Eq. (5), analogous to the notation in Harrer et al. (2021). In this equation, $\widehat{DR_k}$ is the observed discrimination ratio of each correspondence experiment k, $\widehat{DR}$ is the estimated pooled discrimination ratio, $\widehat{\beta}$ is the coefficient representing the fixed effect, x is the study-level variable, p stands for the number of predictors, $\widehat{\varepsilon}$ symbolises the sampling error, and $\widehat{\zeta}$ is the error related to the overarching distribution of true discrimination ratios representing the random effect (also see Section 2.4.1). In the reported models, standard errors related to the model coefficients were clustered at the study level (Viechtbauer, 2010).

$$\widehat{DR_k} = \widehat{DR} + \widehat{\beta}_1 x_{1k} + \ldots + \widehat{\beta}_p x_{pk} + \widehat{\epsilon}_k + \widehat{\zeta}_k \tag{5}$$

2.4.3. Publication bias

There are several reasons why publication bias could adversely impact the results from our meta-analysis, resulting in either an over- or underestimation of hiring discrimination in certain cases. For example, some studies could have been withheld from publication because the results were non-significant, uninteresting or inconclusive while other studies with interesting, significant, or substantial effects were not (i.e. outcome reporting bias). There is also a risk of language bias as only studies in English were included in the review (i.e. language bias). Furthermore, the time lag between the conducting of the research and the publishing of the research could make some studies stay under the radar (i.e. time-lag bias; see Section 3.1).

To measure and counter the impact of publication bias, we applied three analytic techniques: (i) a graphical inspection of funnel plot asymmetry, (ii) the calculation of a 'bias statistic' of funnel plot asymmetry, and (iii) the calculation of bias-adjusted hiring discrimination estimates through 'limit' meta-analyses (Harrer et al., 2021). First of all, we constructed contour-enhanced funnel plots setting off the discrimination ratio of the correspondence experiments against their standard errors. These plots were overlaid by multiple funnel shapes depicting (i) the 95% and 99% confidence intervals around the estimated pooled discrimination ratio for a given discrimination ground or treatment group wherein the observations are expected to fall and (ii) the 90%, 95%, and 99% confidence intervals around the null effect (or discrimination ratio of 1) for a given discrimination ground or treatment group (see Figure A2–1 to Figure A2–16). The second set of funnel shapes helped us in distinguishing outcome reporting bias from other types of publication bias because we could evaluate whether there might be an underreporting of null results.

Second, we used Peters' (2006) binary-effects adaptation of Egger's regression test to calculate a 'bias statistic' for assessing funnel plot asymmetry, which formally compares the discrimination ratios of the respective studies against their standard errors—its null hypothesis assumes that there is no asymmetry. In line with Harrer et al. (2021) and Sterne et al. (2011), we only calculated bias statistics if the total number of correspondence audits for a given analysis equalled or exceeded ten, otherwise, the statistical power could be too low to detect asymmetry.

Third, we replaced some of the original estimates with bias-adjusted estimates obtained from 'limit' meta-analyses. Through these analyses, we allowed for interactions between the observed effects, on the one hand, and the standard error of the pooled effect and the between-study variance, on the other hand (Rücker et al., 2011). This analytical approach resulted in so-called 'shrunken' discrimination ratios that largely account for small-study publication bias (Harrer et al., 2021; Schwarzer et al., 2020). Smaller-sized correspondence experiments are, on average, at greater risk of only being reported if they produce large, statistically significant effects vis-à-vis experiments with larger sample sizes (Borenstein et al., 2009; Harrer et al., 2021). Small-study effects can thus be a source of publication bias due to the correlation between a study's publication status and the nature of its findings. As a general rule, we only calculated these bias-adjusted discrimination ratios (and replaced the original estimates with these ratios) if the total number of correspondence audits for a given analysis equalled or exceeded ten (Harrer et al., 2021; Schwarzer et al., 2015).[14] If not, of the subset

[13] Higgins and Thompson's (2002) guidelines state that values around 25%, 50%, and 75% indicate low, moderate, and high heterogeneity, respectively.

[14] In general, we want to caution the reader when interpreting the pooled estimates for which the included number of correspondence experiments is lower than ten. The uncertainty around the pooled estimates based on a small number of studies can become large, especially in cases where the estimate is based on very few experiments (e.g. only two or three).

of meta-analyses containing fewer than ten studies, the statistical heterogeneity might be too high and the number of observations too low to meaningfully interpret the bias-adjusted discrimination ratios. In cases where we did not calculate bias-adjusted discrimination ratios, we reported the original estimates.[15] The statistical significance of the differences between the original discrimination ratios and the robust, bias-adjusted versions of the discrimination ratios was assessed using *z*-tests (see Table A11; for details on the computational approach, see Altman & Bland, 2003).

2.4.4. Robustness analyses

To further assess the robustness of our results, we (i) measured and controlled for outliers (i.e. influential cases that substantially affect the pooled discrimination ratio), (ii) evaluated potential statistical dependence between the sampled discrimination ratios, and (iii) recalculated *p*-values of the meta-regression coefficients via a permutation test (Borenstein et al., 2009; Higgins et al., 2019; Viechtbauer et al., 2015). First, we identified outliers by looking at studies with extremely small and large discrimination ratios. This process is relevant because it gives us an indication of whether removing these extreme cases from the analyses results in distinctly different discrimination ratios and thus how robust the original estimates are controlling for these outliers. Pooled ratios that are based on only a few correspondence experiments (i.e. units of observation) might be particularly impacted by outliers. In line with Harrer et al. (2021), we defined said ratios as those for which the upper (lower) bound of the 95% confidence interval was lower (higher) than the lower (upper) bound of the confidence interval of the pooled discrimination ratio. To clarify, this means that the ratios of these influential cases were so extreme that they significantly differed from the pooled ratio (at the 5% significance level). Eventually, we recalculated the pooled discrimination ratios, excluding these outliers, and evaluated whether they significantly differed from the original estimates.

Second, we examined the statistical independence of the sampled discrimination ratios. Interdependency between the discrimination ratios could arise in cases wherein different ratios relied on observations from the same control group (Higgins et al., 2019). For example, if a given experiment consisted of an unmatched design with two distinct treatment groups A and B and one control group C, both $\widehat{DR}_{A-C}$ (i.e. the discrimination ratio comparing A with C) and $\widehat{DR}_{B-C}$ (i.e. the discrimination ratio comparing B with C) would be partly based on identical information related to the same control group. This factor could lead to the underestimation of between-study variability, which could, in turn, result in false-positive pooled discrimination ratios. To examine potential statistical independence, we fitted three-level mixed models including estimates of between-study and within-study heterogeneity as well as two-level models that only included estimates of within-study heterogeneity per treatment group and compared these models using ANOVA (for the computational approach, see Harrer et al., 2021). We found no evidence that the three-level models had a better fit with the data than the two-level models (see Table A10). We can thus assume that our results were not significantly impacted by interdependency between the sampled discrimination ratios.

Third, on each estimated meta-regression model, we performed a permutation test with 1000 iterations to control for possible model overfitting (Harrer et al., 2021; Viechtbauer et al., 2015). This test boils down to iteratively redrawing samples from the underlying data, reordering the data into different permutations, to recalculate the *p*-values associated with the test statistics related to the model coefficients. In essence, this test enabled us to better evaluate whether the coefficients identify true patterns in the data or whether they model statistical noise (Harrer et al., 2021). The output from these permutation tests did not significantly alter the interpretation of our meta-regression results.

3. Results

In Section 3.1, we provide some descriptive statistics regarding the correspondence experiments included in our meta-analysis. Subsequently, in Sections 3.2–3.4, we concentrate on the meta-analytic statistics: (i) the pooled discrimination ratios by discrimination ground, (ii) the heterogeneity of these ratios by treatment group, and (iii) their heterogeneity by call-back classification, region, and period. Where appropriate and relevant, the statistical heterogeneity (i.e. statistical measures quantifying between-study variability) and the robustness of the results are discussed. In Section 3.5, finally, we comment on the potential impact of publication bias. The quasi-exhaustive register of correspondence experiments published between 2005 and 2020 on which our analyses were based can be retrieved in full from Table B1 in Appendix B.

3.1. Descriptive statistics

Fig. 2 shows an increase in the annual number of studies based on the correspondence testing method published between 2005 and 2020. More specifically, the number of ended experiments rises as of 2005—right after the publication of Bertrand and Mullainathan's (2004) study—and continues to increase steadily in subsequent years. There is a remarkable peak in the number of publications in 2019. We see two reasons for this sharp increase: (i) many correspondence experiments that ended in previous years (as early as 2013, but mostly in 2016 and 2017) were not published until 2019, and (ii) the *Journal of Ethnic and Migration Studies* compiled a special issue on ethnic discrimination in the labour market that was first published online in 2019. Logically, there is a lag between the year an experiment ends and the year the study is published. On average, this lag is 2.82 years ($SD = 2.06$). While we used the so-called 'Year

[15] In previous versions of this paper (e.g. the IZA discussion paper; Lippens et al., 2021), the bias-adjusted estimates were only reported in the appendix and the reporting of the results was entirely based on the unadjusted estimates.

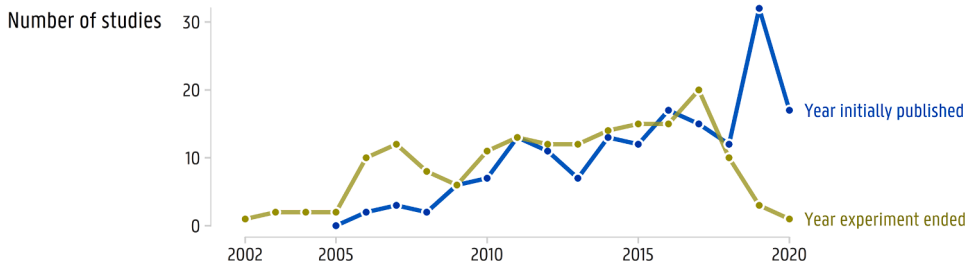

Fig. 2. Time trend of the number of studies based upon correspondence experiments. Notes. 'Year initially published' is the year in which the study was first published (as a pre-print, early-access article, or a full journal article). This year is used in our research as a criterion for study selection, while 'Year experiment ended' is used in our heterogeneity analyses as the period variable.

initially published' as the time-related eligibility criterion in our study selection, the 'Year experiment ended' is used in further heterogeneity analyses because it constitutes a more accurate representation of the timing of a correspondence experiment (see Section 3.4.3).

In our meta-analysis, we focus on two other grouping variables besides time, namely region and call-back classification (see Section 2.4.2). Fig. 3 represents the number of correspondence experiments (i.e. units of observation) by region. The bulk of correspondence experiments is conducted in Europe ($N = 196$, 64.05%), of which 95 are in Western Europe and 60 in Northern Europe, and the Americas ($N = 75$, 24.51%), of which 64 are in Northern America (i.e. mostly the United States). Fig. 4 shows the number of correspondence audits by call-back classification. In the majority of correspondence experiments ($N = 205$, 66.99%), the authors report call-backs in the 'narrow' sense (i.e. an invitation to interview), while call-backs in the 'broad' sense (i.e. any positive response from the employer, such as a request for additional information) are reported in 101 experiments (33.01%). A detailed overview of frequencies and proportions by treatment group and region can be found in Appendix A (Table A2–A3).

Fig. 5 illustrates that the majority of experiments provide results related to the discrimination grounds of race, ethnicity, and national origin ($N = 143$, 46.73%) and gender and motherhood status ($N = 72$, 23.53%). Moreover, relying on counts, there are two discernible patterns concerning the overall treatment effect. First, for most discrimination grounds, there seems to be unequal treatment of applicants from the minority (treatment) group when compared with their majority counterparts. Second, the overall treatment of female gender applicants (vis-à-vis male gender applicants) appears highly ambiguous; in the lion's share of the experiments ($N = 33$, 53.23%), empirical evidence for unequal treatment is absent, while there is hiring discrimination against males and females in the remaining experiments. In Sections 3.2 and 3.3, we meta-analytically assess these treatment effects per discrimination ground and treatment group and address some of the heterogeneity in the uncovered hiring discrimination.

3.2. *Differences in hiring discrimination by discrimination ground*

Table 2 includes the pooled discrimination ratios of the correspondence experiments in our meta-analysis. Unless otherwise indicated, the findings referenced in this section (as well as Sections 3.3 and 3.4) are robust for controlling for outliers.[16,17] Detailed results of these robustness analyses can be found in Appendix A (Table A5–A6 and Table A8–A9). A list of outliers that were removed from the outlier-adjusted statistics can be retrieved from Table A12. In line with the count of votes in Section 3.1, we find empirical evidence for unequal treatment in hiring concerning the discrimination grounds of race, ethnicity, and national origin, age, religion, disability, physical appearance, wealth, and marital status. We also find some evidence of hiring discrimination regarding gender and motherhood status and sexual orientation. However, the uncovered unequal treatment based on gender and motherhood status is very small ($\widehat{DR} = 1.0413$, $CI_{95\%} = [1.0151; 1.0682]$; see also Section 3.3.2) and hiring discrimination related to sexual orientation is not robust when controlling for outliers (*k-adj.* $\widehat{DR} = 0.9007$, $CI_{95\%} = [0.7845; 1.0341]$; see Table A5 and Table A12).[18] Finally, relying on estimates from just four experiments, we find no overall evidence of hiring discrimination based on military service or affiliation ($\widehat{DR} = 0.9983$, $CI_{95\%} = [0.7766; 1.2834]$).

The pooled discrimination ratios enable us to compare the severity of unequal treatment in hiring across different discrimination grounds. Based on these point estimates, people with disabilities are on average approximately 41% less likely to receive a positive response to a job application ($\widehat{DR} = 0.5885$, $CI_{95\%} = [0.5277; 0.6563]$), while estimates based on physical appearance and age indicate

[16] We defined outliers as those discrimination ratios for which the upper (lower) bound of the 95% confidence interval was lower (higher) than the lower (upper) bound of the confidence interval of the pooled discrimination ratio (see section 2.4.4 for more details).

[17] Different from previous versions of this paper (e.g. the IZA discussion paper; Lippens et al., 2021) and as a general rule, if $k \geq 10$, bias-adjusted estimates are reported, otherwise the original estimates are provided (Harrer et al., 2021). These bias-adjusted estimates account for potential small-study publication bias as small-sample studies might bias the original estimates because they are at greater risk of only being reported or published if they produce large, statistically significant effects compared to large-sample studies (see section 2.4.3).

[18] The abbreviation '*k*-adj.' is short for '*k*-adjusted', which indicates that a lower number of studies were included in the analysis, adjusting for influential cases (i.e. outliers). An overview of the specific outliers that were removed (by type of analysis) can be retrieved from Table A12.

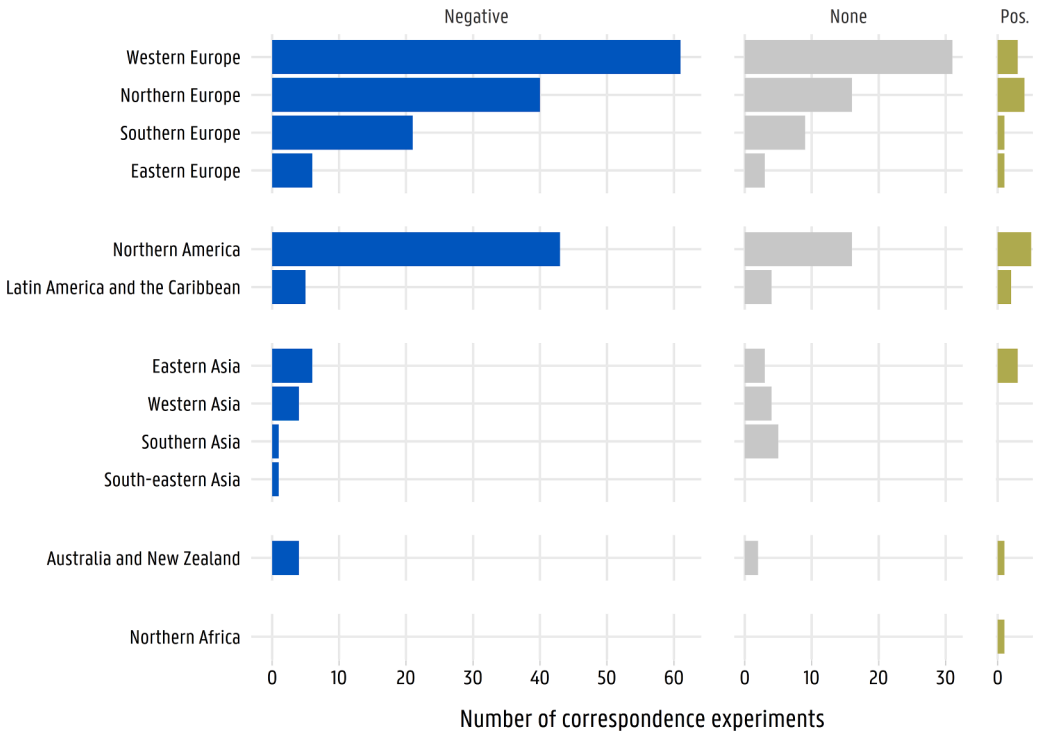

Fig. 3. Number of correspondence experiments by region (rows) and treatment effect (panels). Notes. 'Number of correspondence experiments' represents the units of observation included in the meta-analysis. The bars are grouped in panels, representing the overall treatment effect in the original correspondence experiments. Regional classification is based on the United Nations (2021) M49 Standard. Abbreviations used: Pos. (Positive).

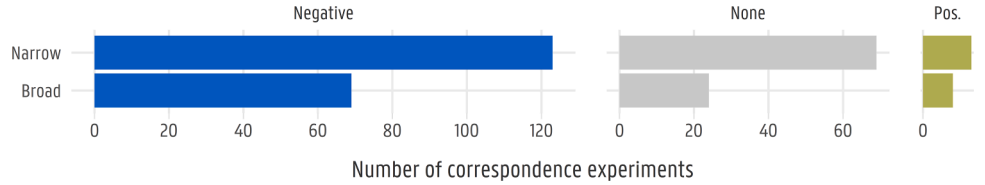

Fig. 4. Number of correspondence experiments by call-back classification (rows) and treatment effect (panels). Notes. Outcome variables consisting of an invitation to a job interview (or any broadly defined positive response from the employer, such as a request for additional information) are classified as narrow (or broad). 'Number of correspondence experiments' represents the units of observation included in the meta-analysis. Bars are grouped in panels, representing the overall treatment effect in the original correspondence experiments. Abbreviations used: Pos. (Positive).

reduced positive responses by approximately 37% ($\widehat{DR} = 0.6308$, $CI_{95\%} = [0.4738; 0.8397]$) and 31% ($\widehat{DR} = 0.6867$, $CI_{95\%} = [0.6503; 0.7250]$), respectively. This contrasts with the discrimination ratios for marital status ($\widehat{DR} = 0.8846$, $CI_{95\%} = [0.8109; 0.9650]$) and wealth ($\widehat{DR} = 0.8806$, $CI_{95\%} = [0.8081; 0.9596]$), which are significantly different from but closer to one. Although we must note that for pooled discrimination ratios that rely on few studies (e.g. $k < 10$), the uncertainty around these estimates can become large and the overall effect ambiguous—we urge caution in interpreting the discrimination ratios in these cases. Most notably, in recent years, many research efforts have focused on examining hiring discrimination based on race, ethnicity, and national origin. Ethnic minority candidates face on average approximately 29% fewer positive responses ($k = 143$, 46.73% of total units of observation; $\widehat{DR} = 0.7113$, $CI_{95\%} = [0.6924; 0.7307]$). Nonetheless, the unequal treatment of disabled, older, and less physically attractive candidates appears at

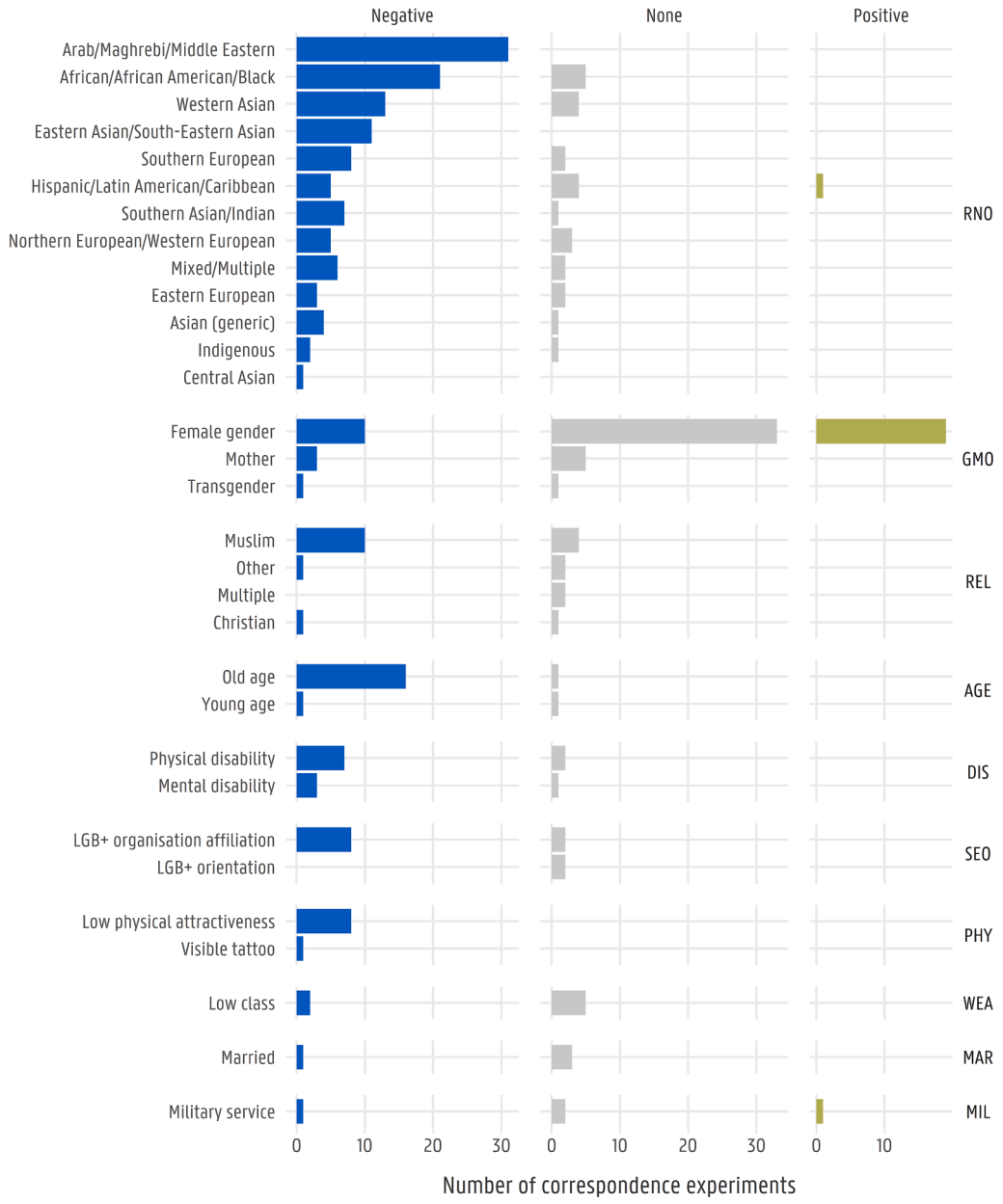

Fig. 5. Number of correspondence experiments by treatment group (rows) and treatment effect (panels). Notes. 'Number of correspondence experiments' represents the units of observation included in the meta-analysis. The bars are grouped in panels, representing the overall treatment effect in the original correspondence experiments. Abbreviations used: RNO (race, ethnicity, and national origin), GMO (gender and motherhood status), AGE (age), REL (religion), DIS (disability), SEO (sexual orientation), PHY (physical appearance), WEA (wealth), MIL (military service or affiliation), and MAR (marital status).

least equally problematic.[19]

In terms of statistical heterogeneity, we witness high variability in the underlying distribution of true discrimination ratios. Specifically, I^2 estimates range from 82.36% (for age) to 98.53% (for sexual orientation)—not considering the exceptional cases of wealth,

[19] As noted by an anonymous reviewer, it is easier to manipulate ethnicity or gender in correspondence experiments because, in many cases, the researcher merely has to change the name on the resume. In contrast, for several other discrimination grounds, the researcher would have to add organisation affiliations (e.g. for sexual orientation or military affiliation), integrate different levels of work experience (e.g. for age) or manipulate photographs (e.g. for physical appearance). This could be one of the reasons why the research focus in recent experimental research on hiring discrimination has been on ethnicity and not on other discrimination grounds.

Table 2
Pooled discrimination ratios by discrimination ground and treatment group.

Variable Discrimination ground or treatment group	Effect k	N_{obs}	N_{events}	$\widehat{DR}$ [$CI_{95\%}$]	t (p)	Statistical heterogeneity I^2
Race, ethnicity, and national origin	143	340,262	68,946	0.7113 [0.6924; 0.7307]	−24.79*** (<0.001)	90.13%
Arab/Maghrebi/Middle Eastern	31	69,311	14,523	0.5937 [0.5548; 0.6353]	−15.09*** (<0.001)	87.37%
African/African American/Black	26	69,177	12,796	0.6845 [0.6444; 0.7270]	−12.32*** (<0.001)	88.44%
Western Asian	17	34,828	7213	0.7508 [0.6977; 0.8080]	−7.66*** (<0.001)	68.13%
Eastern Asian/South-Eastern Asian	11	43,688	5528	0.6286 [0.5368; 0.7361]	−5.77*** (<0.001)	93.06%
Hispanic/Latin American/Caribbean	10	15,344	3632	0.9220 [0.8091; 1.0507]	−1.22 (0.223)	80.52%
Southern European	10	19,648	4036	0.6673 [0.5711; 0.7798]	−5.09*** (<0.001)	76.97%
Mixed/Multiple	8	14,157	3219	0.6757 [0.4287; 1.0651]	−2.04 (0.081)	86.22%
Southern Asian/Indian	8	18,987	4200	0.7004 [0.6352; 0.7723]	−8.61*** (<0.001)	53.16%
Northern European/Western European	8	12,983	3135	0.8154 [0.6661; 0.9981]	−2.39* (0.048)	76.49%
Asian (generic)	5	7098	3114	0.6739 [0.4530; 1.0024]	−2.76 (0.051)	77.77%
Eastern European	5	12,891	2907	0.7206 [0.5271; 0.9851]	−2.91* (0.044)	92.31%
Indigenous	3	20,189	3906	0.7793 [0.4127; 1.4715]	−1.69 (0.233)	95.71%
Central Asian	1	1961	737	N/A	N/A	N/A
Gender and motherhood status	72	330,600	56,650	1.0413 [1.0151; 1.0682]	3.11** (0.002)	94.07%
Female gender	62	308,840	53,309	1.0413 [1.0138; 1.0696]	2.97** (0.003)	94.81%
Mother	8	19,394	2471	0.9044 [0.7887; 1.0370]	−1.74 (0.126)	30.49%
Transgender	2	2366	870	0.8500 [0.5306; 1.3619]	−4.38 (0.143)	N/A
Age	19	86,730	11,775	0.6867 [0.6503; 0.7250]	−13.54*** (<0.001)	82.36%
Old age	17	82,642	11,220	0.6646 [0.6292; 0.7020]	−14.64*** (<0.001)	83.53%
Young age	2	4088	555	0.7698 [0.2294; 2.5830]	−2.75 (0.222)	3.65%
Religion	21	41,917	11,447	0.7855 [0.7457; 0.8274]	−9.11*** (<0.001)	92.45%
Muslim	14	24,344	4947	0.7730 [0.7069; 0.8452]	−5.65*** (<0.001)	85.73%
Other	3	9652	3684	0.8240 [0.3578; 1.8979]	−1.00 (0.423)	91.32%
Christian	2	4642	1375	0.7293 [0.0075; 71.1483]	−0.88 (0.542)	97.75%
Multiple	2	3279	1441	0.9275 [0.7532; 1.1422]	−4.59 (0.137)	N/A
Disability	13	25,232	4530	0.5885 [0.5277; 0.6563]	−9.53*** (<0.001)	96.80%
Physical disability	9	19,694	4191	0.5369 [0.2607; 1.1056]	−1.99 (0.082)	97.83%
Mental disability	4	5538	339	0.6249 [0.4075; 0.9581]	−3.50* (0.039)	38.43%
Sexual orientation[a]	12	41,763	13,442	0.7016 [0.5138; 0.9581]	−2.50* (0.029)	98.53%
LGB+ organisation affiliation[a]	10	38,520	13,018	0.6482 [0.4539; 0.9257]	−2.75* (0.022)	98.78%
LGB+ orientation	2	3243	424	1.0585 [0.5470; 2.0485]	1.09 (0.471)	N/A
Physical appearance	9	50,070	9981	0.6308 [0.4738; 0.8397]	−3.71** (0.006)	97.88%
Wealth	7	11,517	1903	0.8806 [0.8081; 0.9596]	−3.62* (0.011)	N/A
Marital status	4	18,369	2715	0.8846 [0.8109; 0.9650]	−4.49* (0.021)	N/A
Military service or affiliation	4	18,208	1738	0.9983 [0.7766; 1.2834]	−0.02 (0.985)	67.43%

Notes. Abbreviations and notations used: k (number of correspondence experiments), N_{obs} (total number of observations), N_{events} (total number of positive call-backs), $\widehat{DR}$ (pooled discrimination ratio estimate), $CI_{95\%}$ (95% confidence interval), $\widehat{DR}^{*}CI_{95\%}$, LGB+ (lesbian, gay, and bisexual, amongst other sexual orientations), and N/A (not applicable). Pooled discrimination rates are only calculated for discrimination grounds or treatment groups for which $k > 1$. If $k \geq 10$, bias-adjusted estimates from the 'limit' meta-analyses are reported, otherwise the original estimates are provided (see Section 2.4.3). Following Schwarzer et al. (2015), statistical heterogeneity statistics are calculated for those grounds or groups for which $k > 2$. Following Higgins and Thompson (2002), I^2 values around 25%, 50%, or 75% indicate low, moderate, or high heterogeneity, respectively. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

[a] Because the statistical heterogeneity related to this pooled discrimination ratio is extremely high (i.e. I^2 equals approximately 1), resulting in an inaccurate bias-adjusted estimate, we report the unadjusted estimate instead of the publication bias-adjusted estimate obtained from the 'limit' analysis. Here, we thus make an exception to the rule that for analyses where $k \geq 10$ the adjusted estimate is reported

military service or affiliation, and marital status, which are based on a too low number of correspondence experiments to meaningfully interpret the statistical heterogeneity. This signals that the findings of the experiments clustered within the respective discrimination grounds are highly disparate.[20] However, this is not surprising: similar estimates are expected when pooling the discrimination ratios in such broad categories. In Sections 3.3 and 3.4, we assess design-related heterogeneity based on treatment group, call-back classification, region, and period, which helps pinpoint whether this large underlying statistical variability can be (partly) explained by discrepancies based on these variables across study designs.

3.3. Differences in hiring discrimination by treatment group

In this section, we examine differences in hiring discrimination by treatment group. This analysis provides a more granular view of the pooled discrimination ratios described above, as the pooling of said ratios at the level of the discrimination ground substantially

[20] This confirms our priors that (i) there is between-study variation that is driven by differences in the operationalisation of the experimental groups and, more broadly, how the correspondence experiments are designed, and (ii) the random-effects model is an appropriate model to account for such between-study heterogeneity.

masks relevant information about their underlying variability at the level of the treatment group. Estimates by treatment group are also given in Table 2. Moreover, Fig. 6 illustrates visually the relative change in the probability of a positive call-back for the applicants belonging to the respective treatment groups vis-à-vis their counterparts in the control group.

3.3.1. Race, ethnicity, and national origin

First of all, regarding race, ethnicity, and national origin, unfavourable treatment in hiring is highest for applicants belonging to the groups of Arab, Maghrebi, or Middle Eastern ($\widehat{DR} = 0.5937$, $CI_{95\%} = [0.5548; 0.6353]$), Eastern Asian or South-Eastern Asian ($\widehat{DR} = 0.6286$, $CI_{95\%} = [0.5368; 0.7361]$), and Southern European ($\widehat{DR} = 0.6673$, $CI_{95\%} = [0.5711; 0.7798]$) applicants. On average, these applicants face approximately a 41%, 37%, and 33% reduction in the probability of a positive call-back, respectively. We also find overall hiring discrimination against African, African American, or Black ($\widehat{DR} = 0.6845$, $CI_{95\%} = [0.6444; 0.7270]$), Southern Asian or Indian ($\widehat{DR} = 0.7004$, $CI_{95\%} = [0.6352; 0.7723]$, and Western Asian ($\widehat{DR} = 0.7508$, $CI_{95\%} = [0.6977; 0.8080]$) applicants. Furthermore, of all European treatment groups, Southern Europeans experience hiring discrimination to the largest extent, while the discrimination ratios related to the applicants of Eastern European origin ($\widehat{DR} = 0.7206$, $CI_{95\%} = [0.5271; 0.9851]$) or (White) Northern and Western origin ($\widehat{DR} = 0.8154$, $CI_{95\%} = [0.6661; 0.9981]$) are closer to one, meaning that they face less discrimination.[21,22]

Perhaps more surprisingly, at first, we do not find evidence for overall hiring discrimination against Hispanic, Latin American, or Caribbean applicants ($\widehat{DR} = 0.9220$, $CI_{95\%} = [0.8091; 1.0507]$), despite several individual correspondence experiments providing evidence for the unequal treatment of these applicants in hiring (see Fig. 5). However, when the only identified outlier from the analysis of this minority group is excluded, the pooled discrimination ratio becomes statistically significant at the 5% level (*k-adj.* $\widehat{DR} = 0.8175$, $CI_{95\%} = [0.7095; 0.9420]$; see Table A6 and Table A12). This result is in line with previous review studies by Quillian et al. (2017, 2019), who found that discrimination against applicants of Latin American origin seemed to be generally lower than discrimination against applicants belonging to Black, Middle Eastern, North African, or Asian minority groups.

3.3.2. Gender and motherhood status

Next, we take a closer look at hiring discrimination based on gender and motherhood status. We observe a slightly positive discrimination ratio regarding the female gender; there is an average 4.13% higher probability for female gender candidates of receiving a positive response to an application ($\widehat{DR} = 1.0413$, $CI_{95\%} = [1.0138; 1.0696]$). In contrast, we find no evidence of hiring discrimination related to the treatment groups 'transgender' or 'mother'. After excluding outliers from the analysis, the discrimination ratio for female gender (vis-à-vis male gender) applicants remains statistically significant and positive (*k-adj.* $\widehat{DR} = 1.0663$, $CI_{95\%} = [1.0221; 1.1124]$; see Table A6 and Table A12). However, we note that the statistical heterogeneity related to this pooled discrimination ratio is very high ($I^2 = 94.81\%$; see Table 2)—this is exemplified visually in Fig. 6. As pointed out in Section 3.1, we already know that the majority of correspondence experiments on gender discrimination in hiring find null results while the remaining studies find hiring discrimination against male gender candidates but also against female gender candidates (see Fig. 5).

The above findings are in line with a recent large-scale correspondence audit study conducted in the United States (with more than 80,000 applications) in which the authors found that contact rates for male gender and female gender applicants differed significantly between firms: some employers favoured male gender candidates, while others favoured female gender candidates (Kline et al., 2021). Similar findings arise from another recent large-scale correspondence audit study in Australia (with more than 12,000 applications) where males received substantially more (less) positive responses than females in occupations dominated by male (female) gender workers (Adamovic and Leibbrandt, 2022). In our analysis, we do not go further into what could drive this variability. One explanation is that demand-side factors, such as the influence of certain male- or female-orientated job characteristics on the selection criteria used by employers, may lead to gender-based hiring, but also the prototypicality of a candidate for their gender category (Cortina et al., 2021; Yavorsky, 2019). Van Borm and Baert (2022), for example, found women to be perceived as more social and supportive but less assertive or physically strong than men. Applications from female gender candidates for caregiving job positions could consequently be positively received by employers or recruiters, while the opposite might be true for male gender applicants. Another explanation is that the self-selection of members of one gender group into specific sectors or jobs, creating a predominance of that gender in those sectors

[21] Applicants with Albanian-sounding names are discriminated against in Greece and Italy, applicants with Greek names are unfavourably treated in Canada, applicants of Italian origin are discriminated against in Australia and Belgium, and applicants with a Serbian name and appearance experience hiring discrimination in Austria. The control group always consisted of same-country applicants belonging to their region's majority ethnic group.

[22] The Northern and Western European treatment groups comprised minority (majority) applicants of English (Finnish), French (German), German (Irish, Italian, or Russian), and Latvian or Lithuanian (Russian) origin for whom unequal treatment in hiring was assessed. The Eastern European treatment group consisted of minority (majority) applicants of Russian (Finnish), Romanian (Italian), Ukrainian (Russian or Greek), and Polish (Swedish) origin.

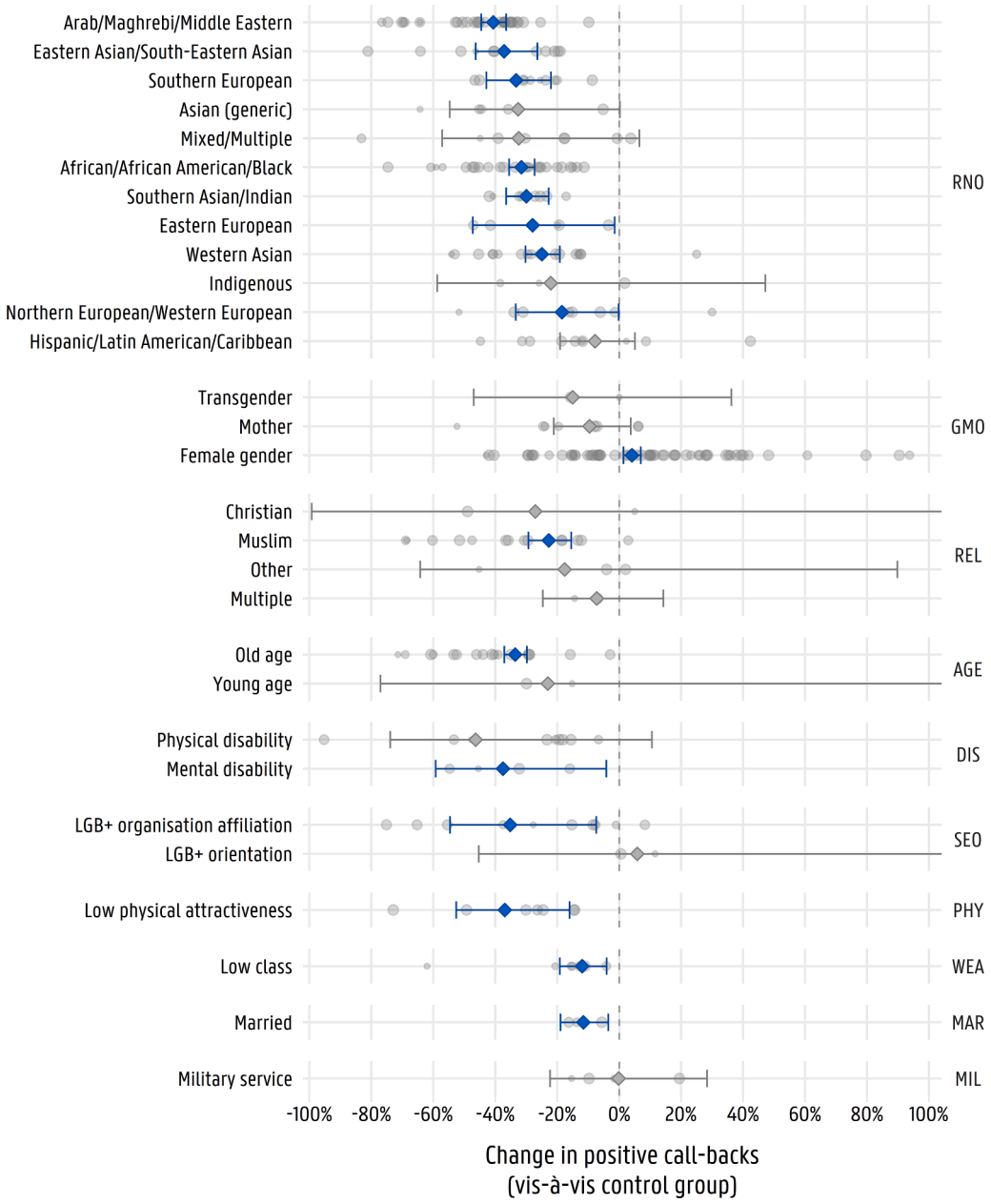

Fig. 6. Hiring discrimination based on pooled discrimination ratios by treatment group. Notes. The change in positive call-backs (compared to the control group), represented by filled diamond shapes, is calculated by subtracting one from the corresponding pooled discrimination ratio (i.e. $\widehat{DR}$; see Table 2). Error bars illustrate the 95% confidence intervals of these ratios. Statistically insignificant ratios (at the 5% level) are greyed out. Semi-transparent dots represent the discrimination ratios of the individual correspondence experiments. Abbreviations used: RNO (race, ethnicity, and national origin), GMO (gender and motherhood status), AGE (age), REL (religion), DIS (disability), SEO (sexual orientation), PHY (physical appearance), WEA (wealth), MAR (marital status), and MIL (military service or affiliation). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

or jobs, could lie at the root of discrimination against members of the opposite group (Cortina et al., 2021).

3.3.3. Religion

The majority of correspondence experiments on religion discrimination in hiring have focused on Muslims ($k = 14$, 66.67% of the total).[23] The remaining correspondence experiments have looked at a highly diverse subset of religions (i.e. evangelical, Jehovah's Witness, Pentecostal, Christian [generic], Buddhist, Hindu, Jewish, no religious affiliation, and various religious affiliations simultaneously). Hiring discrimination based on religion seems to be mainly driven by the unequal treatment of this former group (i.e. Muslims; $\widehat{DR} = 0.7730$, $CI_{95\%} = [0.7069; 0.8452]$). Because of the low density of correspondence experiments concerning other religions, the estimates regarding these treatment groups bear limited reliability. The latter is exemplified by the broad confidence intervals around these estimates (see Table 2 and Fig. 6).

3.3.4. Age

Older applicants (vis-à-vis young to middle-aged applicants), but not younger applicants (vis-à-vis middle-aged applicants), are strongly discriminated against in correspondence experiments ($\widehat{DR} = 0.6646$, $CI_{95\%} = [0.6292; 0.7020]$). Yet, the estimate for younger applicants is based on a very small number of tests ($k = 2$), which warrants caution in interpreting this discrimination ratio. Important to note is that the operationalisation of age differed substantially across correspondence experiments. In this respect, we refer the reader to Appendices A (Table A4) and B (Table B1) for details concerning what constituted 'older' and 'younger' applicants in the original studies. A representative example is the study by Riach (2015), where older candidates were 47 years old, while younger candidates were 27 years old, creating a 20-year age gap between the treatment group and the control group. However, some experiments only look at a 6-year age gap (e.g. Baert et al., 2016), while others go as far as examining a 35-year age gap (e.g. Neumark et al., 2019).

3.3.5. Disability

Discrimination based on disability seems to be equally prompted by the unequal treatment of applicants with a physical disability ($\widehat{DR} = 0.5369$, $CI_{95\%} = [0.2607; 1.1056]$) or a mental disability ($\widehat{DR} = 0.6249$, $CI_{95\%} = [0.4075; 0.9581]$). Although the original estimate related to physical disability is not statistically significant, this discrimination ratio becomes statistically significant at the 5% level after excluding one outlier from the analysis (*k-adj.* $\widehat{DR} = 0.7494$, $CI_{95\%} = [0.6259; 0.8974]$; see Table A6 and Table A12). Here, too, there is a large variety of what is considered a physical disability or mental disability in the original studies (see Table A4 and Table B1). Physical disability includes obesity, blindness or deafness, HIV infection, spinal cord injury, being a wheelchair user, or an unspecified physical disability. Mental disability comprises Asperger's Syndrome, autism, former depression, or a history of mental illness. Given the broad operationalisation of physical disability, it is not surprising that the statistical heterogeneity linked to the pooled discrimination ratio is consequently very high ($I^2 = 97.83\%$; see Table 2 and Fig. 6).

3.3.6. Sexual orientation

Finally, the results concerning hiring discrimination based on sexual orientation are somewhat mixed. The main effect is primarily driven by correspondence experiments considering individuals who have an affiliation with an LGB+ organisation (e.g. membership in an LGB+ rights organisation; $\widehat{DR} = 0.6482$, $CI_{95\%} = [0.4539; 0.9257]$, $k = 10$) in comparison with those who directly disclose an LGB+ orientation ($\widehat{DR} = 1.0585$, $CI_{95\%} = [0.547; 2.0485]$, $k = 2$). Although this finding could raise the question of whether hiring discrimination based on sexual orientation is mainly motivated by a discriminatory stance against activism (i.e. affiliation with an organisation that supports LGB+ rights) rather than discriminatory attitudes regarding LGB+ orientation per se (see also Baert, 2014), the basis for comparison is very narrow. More specifically, we are comparing the results of ten experiments where LGB+ orientation is signalled through an affiliation with those of only two experiments where the LGB+ orientation is disclosed directly.[24] After excluding two clear outliers from the analysis, we also observe that the unequal treatment based on affiliation with an LGB+ organisation becomes statistically insignificant at the 5% level (*k-adj.* $\widehat{DR} = 0.7924$, $CI_{95\%} = [0.6203; 1.0122]$; see Table A6 and Table A12). Overall, the pooled estimates related to sexual orientation seem to be impacted by some form of publication bias (see Table A11 and Figure A2–16).

[23] As pointed out by Bartkoski et al. (2018) and Di Stasio et al. (2021), it is important to make the distinction between hiring discrimination that is purely due to religion (e.g. against Muslims) and hiring discrimination that is due to a combination of origin and religion (e.g. against Arabs). This distinction may have been overseen in previous correspondence experiments. Potentially, in correspondence experiments where the effects of origin and religion interact, confounding might lead to a wrong estimation of the true discrimination. Di Stasio et al. (2021) have attempted to disentangle these effects in their experiment by considering discrimination against 'disclosed Muslims' (i.e. a religion effect) separately from discrimination against 'Muslims by default' (i.e. a religion and/or origin effect).

[24] As an anonymous reviewer has pointed out, other studies focusing on different discrimination grounds could face the same criticism—i.e. the signal used may not be externally valid. For example, in the study of Ameri et al. (2018), disability is signalled in a similar way. Considering race, ethnicity, and national origin, there is an active discussion that African-American or Hispanic names, used to signal African-American or Hispanic status, may also signal socio-economic status (e.g. Darolia et al., 2016; Gaddis, 2017). Moreover, in many of the studies where LGB+ orientation status is signalled through activism or affiliation, there is a similar non-LGB+ signal that is attributed to the control group, which might 'wash out' the activism or affiliation effect (e.g. Drydakis, 2009; Tilcsik, 2011).

Table 3
Differences in pooled discrimination ratios by call-back classification, region, and period per discrimination ground.

Variable			Effect					Statistical heterogeneity
Discrimination ground	Sub-group	Level	k	N_{obs}	N_{events}	$\widehat{DR}$ [$CI_{95\%}$]	t (p)	I^2
Race, ethnicity, and national origin	Call-back	Narrow	87	169,762	35,019	0.6953 [0.6694; 0.7222]	−18.76*** (<0.001)	86.97%
		Broad	56	170,500	33,927	0.7865 [0.7561; 0.8181]	−11.95*** (<0.001)	92.43%
	Region	Americas	38	114,102	17,276	0.8027 [0.7618; 0.8457]	−8.24*** (<0.001)	88.60%
		Europe	94	175,525	46,261	0.7004 [0.6781; 0.7234]	−21.56*** (<0.001)	89.34%
		Asia	6	40,796	3130	0.4877 [0.2900; 0.8203]	−3.55* (0.016)	93.03%
		Other	5	9839	2279	0.6275 [0.4115; 0.9570]	−3.07* (0.037)	90.95%
	Period	2002–2010	45	80,977	18,016	0.6939 [0.6558; 0.7344]	−12.65*** (<0.001)	78.42%
		2011–2020	98	259,285	50,930	0.7211 [0.6992; 0.7436]	−20.81*** (<0.001)	91.57%
Gender and motherhood status	Call-back	Narrow	52	192,641	36,247	1.0766 [1.0449; 1.1092]	4.84*** (<0.001)	87.70%
		Broad	20	137,959	20,403	0.9335 [0.8877; 0.9817]	−2.68** (0.007)	96.87%
	Region	Americas	10	93,335	13,782	1.0498 [0.9971; 1.1053]	1.85 (0.064)	98.20%
		Europe	48	153,130	29,199	1.0190 [0.9840; 1.0552]	1.06 (0.291)	89.92%
		Asia	11	76,260	11,495	1.0422 [0.9795; 1.1089]	1.31 (0.191)	80.37%
		Other	3	7875	2174	1.2854 [0.8355; 1.9775]	2.51 (0.129)	46.28%
	Period	2002–2010	23	93,799	14,824	1.1440 [1.0647; 1.2293]	3.67*** (<0.001)	82.53%
		2011–2020	49	236,801	41,826	0.9878 [0.9612; 1.0151]	−0.88 (0.377)	95.07%
Age	Call-back	Narrow	14	34,541	3549	0.7160 [0.6118; 0.8380]	−4.16*** (<0.001)	82.52%
		Broad	5	52,189	8226	0.6248 [0.5038; 0.7749]	−6.07** (0.004)	74.04%
	Region	Americas	6	61,411	8581	0.6881 [0.6438; 0.7354]	−14.44*** (<0.001)	29.92%
		Europe	13	25,319	3194	0.6288 [0.5349; 0.7392]	−5.62*** (<0.001)	78.87%
		Asia	N/A	N/A	N/A	N/A	N/A	N/A
		Other	N/A	N/A	N/A	N/A	N/A	N/A
	Period	2002–2010	6	19,980	1388	0.5460 [0.4149; 0.7185]	−5.67** (0.002)	52.05%
		2011–2020	13	66,750	10,387	0.6828 [0.6439; 0.7240]	−12.75*** (<0.001)	86.45%
Religion	Call-back	Narrow	11	19,002	4158	0.8350 [0.7484; 0.9317]	−3.23** (0.001)	91.71%
		Broad	10	22,915	7289	0.7425 [0.6984; 0.7894]	−9.53*** (<0.001)	92.91%
	Region	Americas	5	9494	1409	0.7535 [0.5108; 1.1117]	−2.02 (0.113)	81.97%
		Europe	14	26,957	9785	0.8045 [0.7588; 0.8530]	−7.29*** (<0.001)	94.08%
		Asia	2	5466	253	0.6336 [0.0015; 265.6794]	−0.96 (0.513)	92.74%
		Other	N/A	N/A	N/A	N/A	N/A	N/A
	Period	2002–2010	4	7273	1489	0.5633 [0.2738; 1.1588]	−2.53 (0.085)	77.96%
		2011–2020	17	34,644	9958	0.8199 [0.7756; 0.8668]	−7.00*** (<0.001)	90.16%
Disability		Narrow	11	22,645	3580			97.30%

(*continued on next page*)

Table 3 (*continued*)

Variable			Effect					Statistical heterogeneity
Discrimination ground	Sub-group	Level	k	N_{obs}	N_{events}	$\widehat{DR}$ [$CI_{95\%}$]	t (p)	I^2
	Call-back					0.5332 [0.4643; 0.6123]	−8.91*** (<0.001)	
		Broad	2	2587	950	0.7586 [0.4853; 1.1859]	−7.86 (0.081)	N/A
	Region	Americas	5	13,382	1198	0.7308 [0.5198; 1.0275]	−2.56 (0.063)	70.41%
		Europe	8	11,850	3332	0.4722 [0.2110; 1.0569]	−2.20 (0.064)	98.02%
		Asia	N/A	N/A	N/A	N/A	N/A	N/A
		Other	N/A	N/A	N/A	N/A	N/A	N/A
	Period	2002–2010	3	7494	2035	0.3159 [0.0054; 18.5733]	−1.22 (0.348)	99.40%
		2011–2020	10	17,738	2495	0.8017 [0.7080; 0.9077]	−3.49*** (<0.001)	61.76%
Sexual orientation[a]	Call-back	Narrow	10	35,807	12,407	0.6534 [0.4540; 0.9404]	−2.64* (0.027)	98.77%
		Broad	2	5956	1035	0.9970 [0.3606; 2.7565]	−0.04 (0.976)	51.61%
	Region	Americas	3	7179	777	0.7650 [0.4181; 1.4000]	−1.91 (0.197)	75.81%
		Europe	8	30,058	11,220	0.7735 [0.5378; 1.1126]	−1.67 (0.139)	97.78%
		Asia	1	4526	1445	0.2489 [0.2218; 0.2793]	−23.67*** (<0.001)	N/A
		Other	N/A	N/A	N/A	N/A	N/A	N/A
	Period	2002–2010	5	17,678	3719	0.6159 [0.3468; 1.0936]	−2.34 (0.079)	97.84%
		2011–2020	7	24,085	9723	0.7718 [0.4758; 1.2520]	−1.31 (0.238)	98.73%

Notes. Abbreviations and notations used: k (number of correspondence experiments), N_{obs} (total number of observations), N_{events} (total number of positive call-backs), $\widehat{DR}$ (pooled discrimination ratio estimate), $CI_{95\%}$ (95% confidence interval), $\widehat{DR}^{*}$ $CI_{95\%}$, and N/A (not applicable). 'Narrow' refers to correspondence experiments in which the call-back variable is related to an invitation to a job interview; 'broad' refers to experiments in which said variable conveys any positive reaction to an application (e.g. an employer's request for additional. Discrimination rates at the sub-group level are only calculated for discrimination grounds for which the combined number of correspondence experiments equals or exceeds 10. If $k \geq 10$ for a given sub-group level, bias-adjusted estimates from the 'limit' meta-analyses are reported, otherwise the original estimates are provided (see Section 2.4.3). Following Schwarzer et al. (2015), statistical heterogeneity statistics are only calculated for those grounds for which $k > 2$. Following Higgins and Thompson (2002), I^2 values around 25%, 50%, or 75% indicate low, moderate, or high heterogeneity, respectively. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

[a] Because the statistical heterogeneity related to these pooled discrimination ratios is extremely high (i.e. I^2 equals approximately 1 for some estimates), resulting in an inaccurate bias-adjusted estimate, we report the unadjusted estimate instead of the publication bias-adjusted estimate obtained from the limit analysis. Here, we thus make an exception to the rule that for analyses where $k \geq 10$ the adjusted estimate is reported

3.4. Differences in hiring discrimination by call-back classification, region, and period

In this section, we report on the heterogeneity of our results by call-back classification, region, and period. Table 3 contains the pooled discrimination ratios by sub-group and sub-group level per discrimination ground. These ratios are discussed alongside those of the heterogeneity analyses by treatment group, which are included in Appendix A (Table A7). Moreover, Table 4 contains the results from the weighted least squares meta-regression of hiring discrimination on call-back classification, region, and period per discrimination ground. Analogously, the results from similar meta-regression analyses at the level of the treatment group can also be retrieved from Appendix A (Table A14–A15).

3.4.1. Call-back classification heterogeneity

In the hiring discrimination literature, the classification of the outcome measure (i.e. call-back) is usually approached in two ways. Either the reported call-back is measured in the broad sense, where any positive response to the application is interpreted as a positive response, or the call-back is measured in the narrow sense, where only an invitation to interview is regarded as a positive response (but sometimes also both). We assume that hiring discrimination may be lower if the call-back is measured in the broad sense compared to the narrow sense. On the one hand, when considering call-back in the broad sense, this measure also includes so-called 'courtesy questions' where the recruiter or employer poses a question to the minority candidate (but not to the majority candidate) without

Table 4
Weighted least squares meta-regression of hiring discrimination on call-back classification, region, and period.

		RNO	GMO	AGE	REL	DIS	SEO
	Intercept	−35.1630* (15.4015)	19.9726 (12.9508)	−19.0039 (25.6796)	−28.7969 (78.6753)	−59.1381 (136.8726)	−180.4927 (93.3485)
Call-back	*Narrow (ref.)*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Broad	−0.0594 (0.0637)	0.0123 (0.0820)	−0.1189 (0.0944)	−0.0370 (0.2555)	0.2147 (0.2538)	0.3013 (0.2043)
Region	*Americas (ref.)*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	Africa	N/A	0.7105*** (0.1268)	N/A	N/A	N/A	N/A
	Asia	−0.3906 (0.2133)	0.0743 (0.1293)	N/A	−0.1604 (0.6872)	N/A	−1.1651* (0.2440)
	Europe	−0.0833 (0.0687)	0.0356 (0.1192)	−0.2839* (0.0974)	−0.1127 (0.2134)	−0.2758 (0.2193)	−0.0775 (0.2763)
	Oceania	−0.0554 (0.3084)	0.1757 (0.1230)	N/A	N/A	N/A	N/A
Period	Year	0.0173* (0.0077)	−0.0099 (0.0065)	0.0093 (0.0128)	0.0142 (0.0391)	0.0292 (0.0679)	0.0896 (0.0464)
Model	k	143	72	19	21	13	12
	τ^2	0.067	0.047	0.028	0.081	0.489	0.044
	I^2	90.45%	92.11%	77.69%	93.07%	96.23%	89.41%
	AIC	81.942	18.702	9.587	25.831	38.128	10.737
	Pseudo-R^2	13.67%	7.93%	42.58%	7.80%	10.98%	79.53%

Notes. Abbreviations and notations used: RNO (race, ethnicity, and national origin), GMO (gender and motherhood status), AGE (age), REL (religion), DIS (disability), SEO (sexual orientation), ref. (reference group), N/A (not applicable or not available), k (number of correspondence experiments or effects), *AIC* (Akaike information criterion). Following Schwarzer et al. (2015), meta-regression analyses were only performed for discrimination grounds for which $k \geq 10$. Presented statistics are coefficient estimates with standard errors between parentheses. Standard errors were clustered at the study level. Negative (positive) coefficients signify more (less) hiring discrimination against minority candidates for a given variable or dummy category. Following Higgins and Thompson (2002), I^2 values around 25%, 50%, or 75% indicate low, moderate, or high heterogeneity, respectively. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

necessarily having the intention of wanting to know the answer to that question (e.g. 'Could you elaborate a bit on your work experience?'). In these cases, it merely serves as an excuse to dismiss the minority candidate, which might, in turn, obfuscate actual discrimination. On the other hand, this assumption aligns with the idea that—based on statistical discrimination theory—the suitability of minority candidates is literally more likely to be questioned (Lippens et al., 2022). Asking questions to the candidate directly is an evident and cost-effective way to (in)validate a recruiter's or employer's possibly stereotypical image of this candidate.

Nevertheless, we find no evidence for differences in hiring discrimination by call-back classification when controlling for region and period effects (see Table 4 and Table A14). The broadness in measuring or reporting call-backs does not seem to relate to the ratio between the probability of a positive call-back for minority applicants and the probability of a positive response for majority-group candidates (i.e. the discrimination ratio). In other words, levels of hiring discrimination do not appear significantly different if the authors measure and record call-backs in the narrow sense or the broad sense.

3.4.2. Region heterogeneity

There are numerous reasons why hiring discrimination may vary across regions. Differences in terms of legislation, public policies, and socio-economic contexts, amongst other things, can lead to differences in the treatment of minority candidates. A first example, associated with race, ethnicity, and national origin, is that European countries have generally known a much larger influx of migrants from North and Sub-Saharan Africa than North America, while the opposite is true for migrants from Central America and Southern, Eastern, and South-Eastern Asia (Abel and Sander, 2014). Depending on the prevailing theoretical frame of reference, this can be (dis)advantageous to the migrated group: a large migration flow of a specific minority group can help in updating stereotypical ideas of individuals from this group (i.e. statistical discrimination) but it can also elicit prejudice against members of this out-group (i.e. taste-based discrimination; Lippens et al., 2022). A second example, related to age, is that the legal framework surrounding pensions and retirement is much stricter in European countries than it is in the United States, which may have a strong effect on social norms regarding the employment of older candidates (Lahey, 2010). Specifically, mandatory retirement ages—effective in many European countries—may signal to employers that it is appropriate for older people not to be active in the labour market anymore when they approach a certain age, inducing age discrimination in hiring.

We find several regional differences in hiring discrimination. The unfavourable treatment of ethnic minority candidates appears to be greater in Asia ($\widehat{DR} = 0.4877$, $CI_{95\%} = [0.2900; 0.8203]$) than in the Americas ($\widehat{DR} = 0.8027$, $CI_{95\%} = [0.7618; 0.8457]$) or Europe ($\widehat{DR} = 0.7004$, $CI_{95\%} = [0.6781; 0.7234]$; see Table 3). However, controlling for call-back classification and period effects, this difference is not statistically significant (see Table 4 and Table A13). Zooming in on the lower-level treatment groups, we do observe that applicants of Western Asian origin (e.g. Azeri, Armenians, Kurds, Uyghurs) experience significantly more discrimination in Asia ($\widehat{DR} = 0.5321$, $CI_{95\%} = [0.4250; 0.6661]$) than Europe ($\widehat{DR} = 0.8008$, $CI_{95\%} = [0.7427; 0.8634]$; $\widehat{\beta} = -0.3561$, $p = 0.025$; Table A7 and Table A15).[25] The former region comprises both Eastern and Western Asian countries (i.e. China, Georgia, and Turkey). These higher levels of hiring discrimination could be explained by (i) the relatively large local presence of these minority groups in Asian labour markets compared to European labour markets and (ii) the negative connotations associated with these groups within these particular regions, which do not necessarily exist in European countries (Asali et al., 2018; Maurer-Fazio, 2013).

Although absolute levels of hiring discrimination based on race, ethnicity, and national origin appear to be higher in Europe ($\widehat{DR} = 0.7004$, $CI_{95\%} = [0.6781; 0.7234]$) than in the Americas ($\widehat{DR} = 0.8027$, $CI_{95\%} = [0.7618; 0.8457]$), we find no statistically significant evidence for such heterogeneity at the (sub-)regional level controlling for call-back classification and period effects ($\widehat{\beta} = -0.0833$; $p = 0.167$; see Table 4). This contrasts with the findings of Quillian et al. (2019) and Zschirnt and Ruedin (2016) who did find higher levels of ethnic hiring discrimination in Europe vis-à-vis the Americas. The result does not change when we adjust the discrimination ratios for outliers—the gap between the estimates even narrows (see Table A8). This discrepancy is presumably because we primarily focused on differences at the (sub-)regional level and not on differences at the country level. When we compare hiring discrimination in several European countries with the United States directly, we observe that the unequal treatment of ethnic minority candidates in hiring is higher in Finland ($k = 3$, $\widehat{\beta} = -0.4890$, $p = <0.001$), France ($k = 7$, $\widehat{\beta} = -0.2720$, $p = 0.043$), and Italy ($k = 6$, $\widehat{\beta} = -0.2754$, $p = 0.023$) but lower in Germany ($k = 10$, $\widehat{\beta} = 0.2378$, $p = 0.033$) after controlling for call-back classification, period, and treatment group effects (see Table A16 and Figure A3-1 to Figure A3-2). This coincides with some of the findings of Quillian et al. (2019). Note that several of these results are based on a small number of correspondence experiments and thus should be interpreted with caution.

Moreover, hiring discrimination based on gender and motherhood status appears lower in Africa than in the Americas ($\widehat{\beta} = 0.7105$, $p = <0.001$; see Table 3). Nevertheless, this result is not generalisable because the coefficient estimate relies on the comparison of just one African correspondence experiment with ten American experiments.

We also observe regional differences in unequal treatment in hiring based on age, controlling for region and period effects. Generally, age discrimination in hiring is more severe in Europe ($\widehat{DR} = 0.6288$, $CI_{95\%} = [0.5349; 0.7392]$) than in the Americas ($\widehat{DR} = 0.6881$, $CI_{95\%} = [0.6438; 0.7354]$; $\widehat{\beta} = -0.2839$, $p = 0.034$; see Table 3 and Table 4).[26] Specifically, older applicants are more severely

[25] We did not identify any correspondence audits originating from the Americas in which applicants of Western Asian origin were considered as the treatment group.

[26] The meta-regression model concerning age accounts for a substantial amount of heterogeneity (Pseudo-R^2 = 42.58% considering the discrimination ground 'age' and Pseudo-R^2 = 57.00% considering the treatment group 'old age'; see Table 4 and Table A14). In other words, the region seems to be an important variable in explaining the variability between correspondence experiments on age discrimination in hiring.

discriminated against in various European countries (i.e. Belgium, France, the United Kingdom, Spain, and Sweden; $\widehat{DR} = 0.5152$, $CI_{95\%} = [0.4258; 0.6234]$) than in the United States ($\widehat{DR} = 0.6916$, $CI_{95\%} = [0.6342; 0.7541]$; $\widehat{\beta} = -0.3423$, $p = 0.010$; see Table A7, Table A14, and Figure A3–3 to Figure A3-4). This finding is exceptional given that the ages in the treatment groups of the European correspondence experiments range from 37 to 56 years, while the ages used in the American studies are generally higher, ranging from 50 to 66 years.[27] Nonetheless, this regional difference is in line with the average employment rate of 55- to 64-year-olds for the period 2002 (the year of the first correspondence experiments regarding age included in this review) to 2017 (the year of the last correspondence experiments regarding age included in this review) in the United States (60.97%) compared with Belgium (36.89%), France (42.32%), and the United Kingdom (59.89%; OECD, 2021). This finding is also in line with the analysis of Lahey (2010) in that the legislation and social norms around working at an advanced age are on average more lenient in the United States compared to European countries.

Last, we initially observe that hiring discrimination against applicants who are affiliated with an LGB+ organisation or who signal to be LGB+ orientated is higher in Asia ($\widehat{DR} = 0.2489$, $CI_{95\%} = [0.2218; 0.2793]$) than in the Americas ($\widehat{DR} = 0.7650$, $CI_{95\%} = [0.4181; 1.400]$; $\widehat{\beta} = -1.1651$, $p = 0.019$) or Europe ($\widehat{DR} = 0.7735$, $CI_{95\%} = [0.5378; 1.1126]$; $\widehat{\beta} = -1.0876$, $p = <0.001$; see Table 3, Table 4, and Table A13). Here, too, we controlled for call-back classification and period effects. However, we have reasons to believe that this is a spurious correlation: the discrimination ratio of the Asian region is based on just one study from Cyprus (Drydakis, 2014), which is part of Western Asia according to the United Nations M49 Standard classification to which we adhered. Moreover, if we only consider applicants who disclose their sexual orientation via the signal of an LGB+ organisation, the regional difference is no longer significant (see Table A14–A15). In contrast with Flage (2020), we thus find no robust evidence that hiring discrimination based on sexual orientation is higher in Europe than in the Americas.

3.4.3. *Period heterogeneity*

There are diverging opinions about the extent to which discrimination has varied over time (Quillian et al., 2017). Recent meta-analytic evidence that relies on similar, causal evidence of hiring discrimination, however, suggested that there are few to no temporal changes in unequal treatment in hiring based on race, ethnicity, and national origin in the United States and the United Kingdom (Heath and Di Stasio, 2019; Quillian et al., 2017). Relying on data from recent correspondence experiments published between 2005 and 2020 (and conducted between 2002 and 2020), we reassess this evidence. At the same time, we also evaluate temporal heterogeneity in hiring discrimination by gender and motherhood status, age, religion, disability, and sexual orientation. Table 4 shows the results from the meta-regression including the time variable. A visual representation of this heterogeneity in hiring discrimination by discrimination ground and, in minor order, by region, sub-region, or country can be found in Appendix A (Figure A1–1 to Figure A1–16).

In contrast with the results of the meta-studies of Heath and Di Stasio (2019) and Quillian et al., (2017), we do initially find an overall decline in ethnic hiring discrimination. The correlation between (i) the weighted call-back ratios of the individual correspondence experiments related to race, ethnicity, and national origin and (ii) the years in which the respective experiments ended is negative and small but statistically significant, even after controlling for call-back classification and region effects ($r = -0.21$, $k = 143$; $\widehat{\beta} = 0.0173$, $p = 0.030$; see Table 4 and Figure A1–1).[28] This equates to an average increase in positive call-backs for ethnic minorities of 15.52 percentage points between 2006 ($\widehat{DR} = 0.6053$) and 2020 ($\widehat{DR} = 0.7605$).

The decline in ethnicity-based hiring discrimination is primarily driven by the moderate negative correlation related to European correspondence experiments ($r = -0.37$, $k = 94$; $\widehat{\beta} = 0.0267$, $p = 0.001$)—an average increase of 24.32 percentage points in positive call-backs for ethnic minorities between 2006 ($\widehat{DR} = 0.5886$) and 2020 ($\widehat{DR} = 0.8318$)—as opposed to studies conducted in the Americas, where no significant temporal change is observed ($r = 0.13$, $k = 38$; $p = 0.422$; see Table A18 and Figure A1–2).[29] Zooming in on the sub-regions, ethnic hiring discrimination seems to have been mainly in decline in Eastern Europe ($r = -0.84$, $k = 8$) and Western Europe ($r = -0.42$, $k = 48$; see Figure A1–3)—although the former finding is based on too few studies to make conclusive claims.

However, the temporal decrease in ethnic hiring discrimination in Europe does not occur to be robust. The decline becomes statistically insignificant when controlling for the considered treatment groups or when restricting the analysis to each country for which we have data separately (see Table A16, Table A18 and Figure A1–4). In other words, the choice of minority groups in correspondence experiments across European countries in combination with the timing of the experiments seems to have played a meaningful role in the declining figures of hiring discrimination based on race, ethnicity, and national origin. More specifically, in the 2005–2014 period, applicants of Northern, Eastern, and Western European or Western Asian origin received proportionately more attention and applicants of Arab/Maghrebi/Middle Eastern origin less than in the 2015–2020 period, while the former applicant groups face less hiring

[27] As a reminder, Table B1 in Appendix B includes details about the ages of the candidates in the treatment and control groups in the related correspondence experiments.

[28] The correlation coefficient was calculated as the weighted correlation between the majority/minority response ratios of the individual correspondence experiments and the year these experiments ended. Similarly, the regression coefficient was derived from the weighted least squares (WLS) model with said response ratios as the dependent variable and the year the experiments ended as the independent variable. We used the majority/minority response ratios because this allows us to interpret a negative (positive) correlation in terms of a decrease (increase) in hiring discrimination. Weights were derived from the meta-analytic random-effects model (see section 2.4.1).

[29] If any effect, there rather seems to be an upward trend in ethnic hiring discrimination in the Americas.

discrimination than the latter group (see Section 3.3.1). Regarding the remaining discrimination grounds (i.e. gender and motherhood status, age, religion, disability, and sexual orientation), we also find no structural, robust evidence for varying levels of hiring discrimination in recent years.[30] Taken together, our general impression is that there is limited change in hiring discrimination across discrimination grounds over time.

3.5. Publication bias

Appendix A (Table A11 and Figure A2–1 to Figure A2–16) contains useful information to evaluate publication bias. Publication bias is a problem related to the in- and exclusion of studies in a meta-analysis, potentially resulting in an over- or underestimation of the pooled estimates (Harrer et al., 2021). Based on the assessment of (i) funnel plot asymmetry, (ii) the related bias statistics, and (ii) statistical differences between the unadjusted and publication bias-adjusted estimates obtained from the 'limit' meta-analyses, we find that there is potential publication bias regarding the discrimination grounds of race, ethnicity, and national origin and sexual orientation—we find no structural evidence for publication bias associated with the other discrimination grounds.[31,32]

As explained in Section 2.4.3, we have attempted to limit the influence of publication bias to a minimum. Where appropriate and relevant, (i) the bias-adjusted estimates were reported instead of the unadjusted estimates, (ii) results were cross-checked with the outlier-adjusted pooled discrimination ratios, and (iii) discrimination ratios and meta-regression models were also calculated at the lower-level treatment groups (instead of only at the level of the discrimination ground), where there is generally less between-study heterogeneity and thus less (influence of) outliers. Overall, and unless reported otherwise in Sections 3.2–3.4, publication bias had a limited impact on the interpretability of the results.

4. Conclusion

In this meta-analysis, we extensively documented and synthesised the recent hiring discrimination literature grounded in the correspondence testing method—i.e. the reference method of measurement that allows for a causal interpretation of the empirical evidence of unequal treatment in hiring. Unique to our study is the focus on differences in hiring discrimination across discrimination grounds. More concretely, based on experiments from around the world, we quantified the level of hiring discrimination for ten grounds based on which unequal treatment is forbidden under United States federal or state law: (i) race, ethnicity, and national origin, (ii) gender and motherhood status, (iii) age, (iv) religion, (v) disability, (vi) sexual orientation, (vii) physical appearance, (viii) wealth, (ix) marital status, and (x) military service or affiliation. Moreover, we assessed the heterogeneity in hiring discrimination according to the classification of the call-back variable, the region linked to the correspondence experiment, and the related period. Our study provides scholars and policymakers with an extensive comparison based on hiring discrimination research from across the world. Knowing which and to what extent minority groups face labour market inaccessibility is invaluable in tackling this issue. In the following paragraphs, we first discuss the most important results of our meta-analysis, followed by the limitations of our research together with some suggestions for future research.

We observe four notable findings from our analyses. Our first observation relates to the results concerning hiring discrimination at the level of the discrimination ground. Historically, research efforts have focused heavily on examining hiring discrimination based on race, ethnicity, and national origin. This research commitment is not unjustified: applicants with salient racial or ethnic characteristics considerably different from those of the respective majority group(s) in a given country are significantly less likely to receive positive responses to their applications. Specifically, ethnic minority candidates on average receive nearly one-third fewer positive responses to their applications than their majority counterparts. However, it appears that the unequal treatment of applicants with disabilities, older applicants, and less physically attractive applicants is equally problematic. Applicants with disabilities or who have an odd physical appearance receive about two-fifths fewer positive responses on average, while the penalty for old(er) applicants is just above one-third. In addition, we found more modest evidence of hiring discrimination based on religion, wealth, and marital status. Diversity policies, such as outreach campaigns and diversity training, as well as other remedial measures, should also focus on these discrimination grounds. Our meta-analysis underlines that 'diversity in the labour market' should also have a diverse interpretation.

Second, levels of hiring discrimination against the specific minority groups within the set of examined discrimination grounds generally differ substantially. For example, candidates of Arab, Maghrebi, or Middle Eastern origin are severely discriminated against in the hiring process, facing an estimated average reduced chance of a positive response of about two-fifths. At the same time, there is only weak evidence of discrimination against (White) European minority applicants. Therefore, in the first place, measures to decrease hiring discrimination should be targeted at those minority groups who are penalised the most. In this respect, our meta-analysis offers an account of the severity of hiring discrimination against a multitude of minority groups.

Third, we found that there is more hiring discrimination against older applicants (vis-à-vis younger applicants) in Europe (i.e.

[30] Based on Figure A1-16 in Appendix A, one might believe that discrimination based on sexual orientation has declined between 2007 and 2013. However, controlling for call-back classification and region effects, no significant effect of period remains (see Table A14–A15). Moreover, there are issues with publication bias—as a result, the findings related to this form of hiring discrimination bear limited generalisability.

[31] As a reminder, following Harrer et al. (2021) and Sterne et al. (2011), publication bias was only assessed for discrimination grounds and treatment groups where $k \geq 10$ (see Table A11). Hence, we did not calculate estimates for the discrimination grounds physical appearance, wealth, military service or affiliation, and marital status.

[32] Looking at the funnel plot regarding the discrimination ground 'disability', there is, however, one clear outlier (see Figure A2-14).

Belgium, France, the United Kingdom, Spain, and Sweden; approximately 50% fewer positive call-backs on average) versus the United States (approximately 30% fewer positive call-backs on average). This finding is in line with the historic employment rates of 55- to 64-year-olds in the respective countries. Future studies could look into the specific mechanisms that drive these regional differences. European countries or institutions might consequently be able to learn from contextual or policy differences with the United States to determine possible counteracting measures.

Fourth, we observed little differences in hiring discrimination over time. Controlling for call-back classification and region effects, we initially found that hiring discrimination based on race, ethnicity, and national origin had decreased in European correspondence experiments. This decline is primarily driven by audit studies from Western European countries. However, when controlling for the minority groups considered in each experiment, the slope becomes statistically insignificant. The original finding also contrasts with our results at the country level: for each country for which there were sufficient observations to calculate a trend separately, we found no evidence for a significant decline in ethnic hiring discrimination. Moreover, we did not find evidence for structural temporal changes in unequal treatment in hiring related to the remaining discrimination grounds within the scope of this review. Overall, hiring discrimination remains a pervasive issue.

Notwithstanding the important contributions of our review, there are a few limitations concerning the research methods we applied. First of all, our research is based on a synthesis of only correspondence experiments, whereas some earlier meta-studies that focused on specific discrimination grounds have also included in-person audits to paint a broader picture of hiring discrimination (e.g. Quillian et al., 2017). However, as we argued in the introduction, in-person audits face a critical limitation. Behavioural differences between applicants, which are hard to control for, could have an undesirable influence on an employer's assessment in a selection context and therefore muddle the relationship between the individual characteristics of interest (e.g. national origin) and the hiring decision.

Moreover, to some extent, our meta-analysis might suffer from publication bias because we did not consider unpublished manuscripts or non-English research. Nonetheless, we statistically evaluated and attempted to control for said bias. In cases where we suspected publication bias, we reported the results with the necessary caveats but, in general, there were few. Especially the estimates regarding sexual orientation appeared to be influenced by publication bias. In addition, the number of included correspondence experiments regarding marital status, wealth, and military service or affiliation was too small to draw any convincing conclusions. Where the density of evidence is low, more experimentation will be needed before scholars can draw more conclusive inferences.

Finally, we did not explain most of the variability around the hiring discrimination estimates, in part because the scope of our review was already very broad but also because many relevant covariates were not retrievable at the study level across discrimination grounds. The variability that we did explain using meta-regression techniques could not be interpreted causally. For example, it is unclear what exactly drove the regional differences between various European countries and the United States in age discrimination in hiring. Discrepancies in legislation and the resulting social norms might have played a role but we cannot rule out alternative explanations. We advise future studies to investigate the drivers of this variability. A first strategy could be to directly assess context heterogeneity via correspondence experiments by examining the correlation between several vacancy, occupation, organisation, or sector characteristics and unequal treatment in hiring at once, eliminating alternative explanations (see e.g. Kline et al., 2021). A second approach could be to use vignette experiments, randomly assigning participants across context factors to observe how prejudice or different stigmas and stereotypes vary (see e.g. Van Borm and Baert, 2022). A third method that we can think of is to apply appropriate meta-regression techniques to sufficiently specific research problems regarding a more restricted selection of minority groups, assessing the relationship between the available study-level variables and hiring discrimination (see e.g. Quillian et al., 2017). This would enable scholars to more precisely attribute the uncovered variance of the pooled discrimination ratios to relevant factors beyond the contextual heterogeneity in terms of call-back classification, region, and period discussed in this review

Funding

This study was conducted in the context of the EdisTools project. EdisTools is funded by Research Foundation – Flanders (Strategic Basic Research, S004119N).

CRediT authorship contribution statement

Louis Lippens: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Siel Vermeiren:** Investigation, Data curation, Writing – original draft. **Stijn Baert:** Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

There are no relevant financial or non-financial competing interests.

Data Availability

The data used in this study are available at the following URL: https://doi.org/10.34740/kaggle/dsv/4142915.

Acknowledgements

We are grateful to the authors of a handful of the audit studies included in this review for providing us with missing data to supplement our dataset and with feedback on a prior version of this paper. Moreover, we are thankful to Brecht Neyt, the participants of the 26th Spring Meeting of Young Economists, and the participants of the 19th IMISCOE Annual Conference for their helpful comments and suggestions. Last, we want to thank three anonymous reviewers for their feedback, which greatly helped to improve this work

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2022.104315.

References

Abel, G.J., Sander, N., 2014. Quantifying global international migration flows. Science 343 (6178), 1520–1522. https://doi.org/10.1126/science.1248676.
Adamovic, M., 2020. Analyzing discrimination in recruitment: a guide and best practices for resume studies. International Journal of Selection and Assessment 28 (4), 445–464. https://doi.org/10.1111/ijsa.12298.
Adamovic, M., 2022. When ethnic discrimination in recruitment is likely to occur and how to reduce it: applying a contingency perspective to review resume studies. Human Resource Management Review 32 (2), 100832. https://doi.org/10.1016/j.hrmr.2021.100832.
Adamovic, M., Leibbrandt, A., 2022. A large-scale field experiment on occupational gender segregation and hiring discrimination. Industrial Relations: A Journal of Economy and Society. https://doi.org/10.1111/irel.12318. Advance online publication.
Altonji, J.G., & Blank, R.M. (1999). Race and gender in the labor market. In O. C. Ashenfelter & D. Card (eds.), *Handbook of Labor Economics* (Vol. 3, pp. 3143–3259). Elsevier. https://doi.org/10.1016/S1573-4463(99)30039-0.
Altman, D.G., Bland, J.M., 2003. Statistics notes: interaction revisited: the difference between two estimates. BMJ 326 (7382), 219. https://doi.org/10.1136/bmj.326.7382.219.
Ameri, M., Schur, L., Adya, M., Bentley, F. S., McKay, P., & Kruse, D. (2018). The disability employment puzzle: A field experiment on employer hiring behavior. ILR Review, 71(2), 329364. https://doi.org/10.1177/0019793917717474.
Asali, M., Pignatti, N., Skhirtladze, S., 2018. Employment discrimination in a former Soviet Union republic: evidence from a field experiment. J Comp Econ 46 (4), 1294–1309. https://doi.org/10.1016/j.jce.2018.09.001.
Baert, S., 2014. Career lesbians. Getting hired for not having kids? Industrial Relations Journal 45 (6), 543–561. https://doi.org/10.1111/irj.12078.
Baert, S., 2018. Hiring discrimination: an overview of (almost) all correspondence experiments since 2005. In: Gaddis, S.M. (Ed.), Audit studies: Behind the Scenes With theory, method, and Nuance. Springer, pp. 63–77. https://doi.org/10.1007/978-3-319-71153-9_3.
Baert, S., 2021. The iceberg decomposition: a parsimonious way to map the health of labour markets. Econ Anal Policy 69, 350–365. https://doi.org/10.1016/j.eap.2020.12.012.
Baert, S., Norga, J., Thuy, Y., Van Hecke, M., 2016. Getting grey hairs in the labour market. An alternative experiment on age discrimination. J Econ Psychol 57, 86–101. https://doi.org/10.1016/j.joep.2016.10.002.
Balduzzi, S., Rücker, G., Schwarzer, G., 2019. How to perform a meta-analysis with R: a practical tutorial. Evidence Based Mental Health 22 (4), 153–160. https://doi.org/10.1136/ebmental-2019-300117.
Balestra, C., & Fleischer, L. (2018). Diversity statistics in the OECD: how do OECD countries collect data on ethnic, racial and indigenous identity? (Organisation for Economic Cooperation and Development [OECD] Statistics Working Papers No. 2018/09). Organisation for Economic Cooperation and Development. https://doi.org/10.1787/89bae654-en.
Bartkoski, T., Lynch, E., Witt, C., Rudolph, C., 2018. A meta-analysis of hiring discrimination against Muslims and Arabs. Personnel Assessment and Decisions 4 (2), 1–16. https://doi.org/10.25035/pad.2018.02.001.
Batinović, L., Howe, M., Sinclair, S., Carlsson, R., 2022. Ageism in hiring: a systematic review and meta-analysis of age discrimination. PsyArXiv. https://doi.org/10.31234/osf.io/sbzmv.
Beam, E.A., Hyman, J., Theoharides, C., 2020. The relative returns to education, experience, and attractiveness for young workers. Econ Dev Cult Change 68 (2), 391–428. https://doi.org/10.1086/701232.
Bertrand, M., Duflo, E., 2017. Field experiments on discrimination. In: Banerjee, A.V., Duflo, E. (Eds.), Handbook of Economic Field Experiments, 1st ed., 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. American Economic Review 94 (4), 991–1013. https://doi.org/10.1257/0002828042002561.
Blinder, A.S., 1973. Wage discrimination: reduced form and structural estimates. J Hum Resour 8 (4), 436–455. https://doi.org/10.2307/144855.
Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2009. Introduction to Meta-Analysis. John Wiley & Sons. https://doi.org/10.1002/9780470743386.
Borjas, G., 2020. Labor market discrimination. In: Borjas, G. (Ed.), Labor Economics, 8th ed. McGraw-Hill, pp. 299–340.
Briner, R.B., & Denyer, D. (2012). Systematic review and evidence synthesis as a practice and scholarship tool. Oxford Handbooks Online. https://doi.org/10.1093/oxfordhb/9780199763986.013.0007.
Boystock v. Clayton County, 590 U.S. __ (U.S. Sup. Ct. 2020). https://www.supremecourt.gov/opinions/19pdf/17-1618_hfci.pdf.
Carlsson, M., Eriksson, S., 2019. Age discrimination in hiring decisions: evidence from a field experiment in the labor market. Labour Econ 59, 173–183. https://doi.org/10.1016/j.labeco.2019.03.002.
Cochran, W.G., 1954. Some methods for strengthening the common χ^2 tests. Biometrics 10 (4), 417–451. https://doi.org/10.2307/3001616.
Cortina, C., Rodríguez, J., González, M.J., 2021. Mind the job: the role of occupational characteristics in explaining gender discrimination. Soc Indic Res 156 (1), 91–110. https://doi.org/10.1007/s11205-021-02646-2.
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. Appl Econ Lett 23 (12), 853–856. https://doi.org/10.1080/13504851.2015.1114571.
Derous, E., Ryan, A.M., 2019. When your resume is (not) turning you down: modelling ethnic bias in resume screening. Human Resource Management Journal 29 (2), 113–130. https://doi.org/10.1111/1748-8583.12217.
Di Stasio, V., Lancee, B., Veit, S., Yemane, R., 2021. Muslim by default or religious discrimination, Results from a cross-national field experiment on hiring discrimination. J Ethn Migr Stud 47 (6), 1305–1326. https://doi.org/10.1080/1369183x.2019.1622826.
Drydakis, N., 2009. Sexual orientation discrimination in the labour market. Labour Econ 16 (4), 364–372. https://doi.org/10.1016/j.labeco.2008.12.003.
Drydakis, N., 2014. Sexual orientation discrimination in the Cypriot labour market: distastes or uncertainty? Int J Manpow 35 (5), 720–744. https://doi.org/10.1108/ijm-02-2012-0026.

Drydakis, N., 2017. Measuring labour differences between natives, non-natives, and natives with an ethnic-minority background. Econ Lett 161, 27–30. https://doi.org/10.1016/j.econlet.2017.08.031.
Drydakis, N., 2022. Sexual orientation and earnings: a meta-analysis 2012–2020. J Popul Econ 35 (2), 409–440. https://doi.org/10.1007/s00148-021-00862-1.
European Commission. (2021). Guidance note on the collection and use of equality data based on racial or ethnic origin. Publications Office of the European Union. https://ec.europa.eu/info/sites/default/files/guidance_note_on_the_collection_and_use_of_equality_data_based_on_racial_or_ethnic_origin.pdf.
Flage, A., 2020. Discrimination against gays and lesbians in hiring decisions: a meta-analysis. Int J Manpow 41 (6), 671–691. https://doi.org/10.1108/ijm-08-2018-0239.
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. https://doi.org/10.1093/sf/sou111.
Gaddis, S.M., 2017. Racial/Ethnic perceptions from Hispanic names: selecting names to test for discrimination. Socius: Sociological Research for a Dynamic World 3, 1–11. https://doi.org/10.1177/2378023117737193.
Gaddis, S.M., 2018. An Introduction to audit studies in the social sciences. In: Gaddis, S.M. (Ed.), Audit studies: Behind the Scenes With theory, method, and Nuance. Springer, pp. 3–44. https://doi.org/10.1007/978-3-319-71153-9_1.
Gaddis, S.M., Larsen, E., Crabtree, C., & Holbein, J. (2021). Discrimination against Black and Hispanic Americans is highest in hiring and housing contexts: a meta-analysis of correspondence audits. (SSRN Electronic Journal Working Paper No. 3975770). University of Chicago, Becker Friedman Institute for Economics. https://doi.org/10.2139/ssrn.3975770.
Ganty, S. & Benito-Sanchez, J.C. (2021). Expanding the list of protected grounds within anti-discrimination law in the EU. Equinet: European Network of Equality Bodies. https://equineteurope.org/wp-content/uploads/2022/03/Expanding-the-List-of-Grounds-in-Non-discrimination-Law_Equinet-Report.pdf.
Guul, T.S., Villadsen, A.R., Wulff, J.N., 2019. Does good performance reduce bad behavior, Antecedents of ethnic employment discrimination in public organizations. Public Adm Rev 79 (5), 666–674. https://doi.org/10.1111/puar.13094.
Harrer, M., Cuijpers, P., Furukawa, T.A., & Ebert, D.D. (2019). dmetar: Companion R package for the guide 'Doing meta-analysis in R' (Version 0.0.9000) [Computer software]. https://dmetar.protectlab.org.
Harrer, M., Cuijpers, P., Furukawa, T.A., Ebert, D.D., 2021. Doing Meta-Analysis With R: A hands-On Guide, 1st ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781003107347.
Havránek, T., Stanley, T.D., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W.R., Rost, K., Aert, R.C.M., 2020. Reporting guidelines for meta-analysis in economics. J Econ Surv 34 (3), 469–475. https://doi.org/10.1111/joes.12363.
Heath, A.F., Di Stasio, V., 2019. Racial discrimination in Britain, 1969–2017: a meta-analysis of field experiments on racial discrimination in the British labour market. Br J Sociol 70 (5), 1774–1798. https://doi.org/10.1111/1468-4446.12676.
Higgins, J.P.T., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Stat Med 21 (11), 1539–1558. https://doi.org/10.1002/sim.1186.
Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A., 2019. Cochrane Handbook For Systematic Reviews of Interventions, 1st ed. Wiley. https://doi.org/10.1002/9781119536604.
Hoaglin, D.C., 2016. Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. Stat Med 35 (4), 485–495. https://doi.org/10.1002/sim.6632.
Jacquemet, N., Yannelis, C., 2012. Indiscriminate discrimination: a correspondence test for ethnic homophily in the Chicago labor market. Labour Econ 19 (6), 824–832. https://doi.org/10.1016/j.labeco.2012.08.004.
Jowell, R., Prescott-Clarke, P., 1970. Racial discrimination and white-collar workers in Britain. Race 11 (4), 397–417. https://doi.org/10.1177/030639687001100401.
Kitagawa, E.M., 1955. Components of a difference between two rates. J Am Stat Assoc 50 (272), 1168. https://doi.org/10.2307/2281213.
Kline, P., Rose, E., & Walters, C. (2021). Systemic discrimination among large US employers. (SSRN Electronic Journal Working Paper No. 2021–94). University of Chicago, Becker Friedman Institute for Economics. https://doi.org/10.2139/ssrn.3898669.
Knapp, G., Hartung, J., 2003. Improved tests for a random effects meta-regression with a single covariate. Stat Med 22 (17), 2693–2710. https://doi.org/10.1002/sim.1482.
Lahey, J.N., 2010. International comparison of age discrimination laws. Res Aging 32 (6), 679–697. https://doi.org/10.1177/0164027510379348.
Lang, K., Kahn-Lang-Spitzer, A., 2020. Race discrimination: an economic perspective. Journal of Economic Perspectives 34 (2), 68–89. https://doi.org/10.1257/jep.34.2.68.
Langan, D., Higgins, J.P.T., Jackson, D., Bowden, J., Veroniki, A.A., Kontopantelis, E., Viechtbauer, W., Simmonds, M., 2019. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Res Synth Methods 10 (1), 83–98. https://doi.org/10.1002/jrsm.1316.
Larsen, E.N., Di Stasio, V., 2021. Pakistani in the UK and Norway: different contexts, similar disadvantage. Results from a comparative field experiment on hiring discrimination. J Ethn Migr Stud 47 (6), 1201–1221. https://doi.org/10.1080/1369183x.2019.1622777.
Lippens, L., Baert, S., Ghekiere, A., Verhaeghe, P.-P., Derous, E., 2022. Is labour market discrimination against ethnic minorities better explained by taste or statistics? A systematic review of the empirical evidence. J Ethn Migr Stud. https://doi.org/10.1080/1369183X.2022.2050191. Advance online publication.
Lippens, L., Vermeiren, S., & Baert, S. (2021). The state of hiring discrimination: a meta-analysis of (almost) all recent correspondence experiments (Institut zur Zukunft der Arbeit [IZA] Discussion Papers No. 14966). IZA Institute of Labor Economics. https://www.iza.org/publications/dp/14966/the-state-of-hiring-discrimination-a-meta-analysis-of-almost-all-recent-correspondence-experiments.
Mantel, N., Haenszel, W., 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 22 (4), 719–748. https://doi.org/10.1093/jnci/22.4.719.
Maurer-Fazio, M., 2013. Ethnic discrimination in China's internet job board labor market. IZA Journal of Migration 1 (12). https://doi.org/10.1186/2193-9039-1-12.
Morning, A., 2008. Ethnic classification in global perspective: a cross-national survey of the 2000 census round. Popul Res Policy Rev 27 (2), 239–272. https://doi.org/10.1007/s11113-007-9062-5.
Neumark, D., 2018. Experimental research on labor market discrimination. J Econ Lit 56 (3), 799–866. https://doi.org/10.1257/jel.20161309.
Neumark, D., Burn, I., Button, P., Chehras, N., 2019. Do state laws protecting older workers from discrimination reduce age discrimination in hiring? Evidence from a field experiment. The Journal of Law and Economics 62 (2), 373–402. https://doi.org/10.1086/704008.
Oaxaca, R., 1973. Male-female wage differentials in urban labor markets. Int Econ Rev (Philadelphia) 14 (3), 693–709. https://doi.org/10.2307/2525981.
Organisation for Economic Cooperation and Development. (2020a). All hands in? Making diversity work for all. https://doi.org/10.1787/efb14583-en.
Organisation for Economic Cooperation and Development. (2020b). International migration outlook 2020. https://doi.org/10.1787/0c0cc42a-en.
Organisation for Economic Cooperation and Development. (2021). Employment rate by age group [Data set]. https://doi.org/10.1787/084f32c7-en.
Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., Moher, D., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, n71. https://doi.org/10.1136/bmj.n71.
Page, M.J., Sterne, J.A.C., Higgins, J.P.T., Egger, M., 2020. Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: a review. Res Synth Methods 12 (2), 248–259. https://doi.org/10.1002/jrsm.1468.
Pager, D., 2016. Are firms that discriminate more likely to go out of business. Sociol Sci 3, 849–859. https://doi.org/10.15195/v3.a36.
Patacchini, E., Ragusa, G., Zenou, Y., 2015. Unexplored dimensions of discrimination in Europe: homosexuality and physical appearance. J Popul Econ 28 (4), 1045–1073. https://doi.org/10.1007/s00148-014-0533-9.
Peters, J.L., 2006. Comparison of two methods to detect publication bias in meta-analysis. JAMA 295 (6), 676–680. https://doi.org/10.1001/jama.295.6.676.
Quillian, L., Midtbøen, A.H., 2021. Comparative perspectives on racial discrimination in hiring: the rise of field experiments. Annu Rev Sociol 47 (1), 391–415. https://doi.org/10.1146/annurev-soc-090420-035144.
Quillian, L., Heath, A., Pager, D., Midtbøen, A., Fleischmann, F., Hexel, O., 2019. Do some countries discriminate more than others? Evidence from 97 field experiments of racial discrimination in hiring. Sociol Sci 6, 467–496. https://doi.org/10.15195/v6.a18.

Quillian, L., Lee, J.J., Oliver, M., 2020. Evidence from field experiments in hiring shows substantial additional racial discrimination after the callback. Social Forces 99 (2), 732–759. https://doi.org/10.1093/sf/soaa026.
Quillian, L., Pager, D., Hexel, O., Midtbøen, A.H., 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. Proceedings of the National Academy of Sciences 114 (41), 10870–10875. https://doi.org/10.1073/pnas.1706255114.
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.
Rich, J. (2014). What do field experiments of discrimination in markets tell us? A meta-analysis of studies conducted since 2000 (IZA Discussion Papers No. 8584). IZA Institute of Labor Economics. https://www.iza.org/publications/dp/8584/what-do-field-experiments-of-discrimination-in-markets-tell-us-a-meta-analysis-of-studies-conducted-since-2000.
Richardson, W.S., Wilson, M.C., Nishikawa, J., Hayward, R.S.A., 1995. The well-built clinical question: a key to evidence-based decisions. ACP J. Club 123 (3), A12. https://doi.org/10.7326/acpjc-1995-123-3-a12.
Rücker, G., Schwarzer, G., Carpenter, J.R., Binder, H., Schumacher, M., 2011. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. Biostatistics 12 (1), 122–142. https://doi.org/10.1093/biostatistics/kxq046.
Rücker, G., Schwarzer, G., Carpenter, J.R., Schumacher, M., 2008. Undue reliance on *I2* in assessing heterogeneity may mislead. BMC Med Res Methodol 8 (1), 79. https://doi.org/10.1186/1471-2288-8-79.
Schwarzer, G., Carpenter, J.R., Rücker, G, 2015. Heterogeneity and meta-regression. In: Schwarzer, G., Carpenter, J.R., Rücker, G. (Eds.), Meta-analysis With R. Springer, pp. 85–104. https://doi.org/10.1007/978-3-319-21416-0_4.
Schwarzer, G., Carpenter, J.R., & Rücker, G. (2020). Metasens: advanced statistical methods to model and adjust for bias in meta-analysis (Version 0.6–0) [Computer software]. https://CRAN.R-project.org/package=metasens.
Sterne, J.A.C., Sutton, A.J., Ioannidis, J.P.A., Terrin, N., Jones, D.R., Lau, J., Carpenter, J., Rucker, G., Harbord, R.M., Schmid, C.H., Tetzlaff, J., Deeks, J.J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D.G., Moher, D., Higgins, J.P.T., 2011. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 343, d4002. https://doi.org/10.1136/bmj.d4002.
Stone, A., Wright, T., 2013. When your face doesn't fit: employment discrimination against people with facial disfigurements. J Appl Soc Psychol 43 (3), 515–526. https://doi.org/10.1111/j.1559-1816.2013.01032.x.
Thijssen, L., Coenders, M., Lancee, B., 2021a. Ethnic discrimination in the Dutch labor market: differences between ethnic minority groups and the role of personal information about job applicants—evidence from a field experiment. J Int Migr Integr 22 (3), 1125–1150. https://doi.org/10.1007/s12134-020-00795-w.
Thijssen, L., van Tubergen, F., Coenders, M., Hellpap, R., Jak, S., 2021b. Discrimination of Black and Muslim minority groups in western societies: evidence from a meta-analysis of field experiments. International Migration Review. https://doi.org/10.1177/01979183211045044. Advance online publication.
Thomas, K., 2018. The labor market value of taste: an experimental study of class bias in US employment. Sociol Sci 5, 562–595. https://doi.org/10.15195/v5.a24.
Tilcsik, A., 2011. Pride and prejudice: employment discrimination against openly gay men in the United States. American Journal of Sociology 117 (2), 586–626. https://doi.org/10.1086/661653.
United Nations. (2021). United Nations standard country codes (Series M: miscellaneous Statistical Papers No. 49). https://unstats.un.org/unsd/methodology/m49/.
Van Borm, H., & Baert, S. (2022). Diving in the minds of recruiters: what triggers gender stereotypes in hiring? (IZA Discussion Papers No. 15261). IZA Institute of Labor Economics. https://www.iza.org/publications/dp/15261/diving-in-the-minds-of-recruiters-what-triggers-gender-stereotypes-in-hiring.
Verhaeghe, P.-P., 2022. Correspondence studies. In: Zimmermann, K.F. (Ed.), Handbook of Labor, Human Resources and Population Economics. Springer, Cham. https://doi.org/10.1007/978-3-319-57365-6_306-1.
Veroniki, A.A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J.P.T., Langan, D., Salanti, G., 2015. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods 7 (1), 55–79. https://doi.org/10.1002/jrsm.1164.
Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. J Stat Softw 36 (3), 1–48. https://doi.org/10.18637/jss.v036.i03.
Viechtbauer, W., López-López, J.A., Sánchez-Meca, J., Marín-Martínez, F., 2015. A comparison of procedures to test for moderators in mixed-effects meta-regression models. Psychol Methods 20 (3), 360–374. https://doi.org/10.1037/met0000023.
Yavorsky, J.E., 2019. Uneven patterns of inequality: an audit analysis of hiring-related practices by gendered and classed contexts. Social Forces 98 (2), 461–492. https://doi.org/10.1093/sf/soy123.
Yemane, R., Fernández-Reino, M., 2021. Latinos in the United States and in Spain: the impact of ethnic group stereotypes on labour market outcomes. J Ethn Migr Stud 47 (6), 1240–1260. https://doi.org/10.1080/1369183x.2019.1622806.
Zschirnt, E., Ruedin, D., 2016. Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests, 1990–2015. J Ethn Migr Stud 42 (7), 1115–1134. https://doi.org/10.1080/1369183X.2015.1133279.