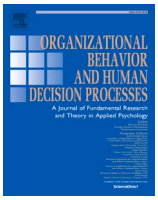
Organizational Behavior and Human Decision Processes 179 (2023) 104280

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

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



On the trajectory of discrimination: A meta-analysis and forecasting survey capturing 44 years of field experiments on gender and hiring decisions

Michael Schaerer [a,1,*], Christilene du Plessis [a,1,*], My Hoang Bao Nguyen [a], Robbie C.M. van Aert [b], Leo Tiokhin [c,j], Daniël Lakens [c], Elena Giulia Clemente [d], Thomas Pfeiffer [e], Anna Dreber [f], Magnus Johannesson [g], Cory J. Clark [h], Gender Audits Forecasting Collaboration[2], Eric Luis Uhlmann [i]

[a] *Singapore Management University, Singapore*
[b] *Tilburg University, The Netherlands*
[c] *Eindhoven University of Technology, The Netherlands*
[d] *University of Zurich, Switzerland*
[e] *Massey University, New Zealand*
[f] *Stockholm School of Economics, Sweden, & University of Innsbruck, Austria*
[g] *Stockholm School of Economics, Sweden*
[h] *University of Pennsylvania, United States*
[i] *INSEAD, Singapore*
[j] *Data & Analytics Group, IG&H, The Netherlands*

ARTICLE INFO

Keywords:
Gender
Discrimination
Field experiments
Meta-analysis
Open science
Forecasting

ABSTRACT

A preregistered meta-analysis, including 244 effect sizes from 85 field audits and 361,645 individual job applications, tested for gender bias in hiring practices in female-stereotypical and gender-balanced as well as male-stereotypical jobs from 1976 to 2020. A "red team" of independent experts was recruited to increase the rigor and robustness of our meta-analytic approach. A forecasting survey further examined whether laypeople (n = 499 nationally representative adults) and scientists (n = 312) could predict the results. Forecasters correctly anticipated reductions in discrimination against female candidates over time. However, both scientists and laypeople overestimated the continuation of bias against female candidates. Instead, selection bias in favor of male over female candidates was eliminated and, if anything, slightly reversed in sign starting in 2009 for mixed-gender and male-stereotypical jobs in our sample. Forecasters further failed to anticipate that discrimination against male candidates for stereotypically female jobs would remain stable across the decades.

Once lay down the rule that the job comes first and you throw that job open to every individual, man or woman, fat or thin, tall or short, ugly or beautiful, who is able to do that job better than the rest of the world.
– Dorothy L. Sayers

1. Introduction

How widespread is gender discrimination in hiring and selection, and have at least some human societies experienced meaningful change towards greater equality of opportunity? These intertwined questions represent two of the most theoretically rich, practically important, and politically controversial scientific issues of our time. For scholars, the answers hold implications for our understanding of the nature of gender stereotypes and the possibility of rapid cultural evolution. For practitioners, they point to different tactics for ensuring the fairness of selection processes into organizations. For citizens and leaders, they may validate or deeply challenge ideological assumptions and worldviews.

Psychological accounts of bias stipulate that group-based discrimination can result from cognitive (Bodenhausen, 1988), motivational

* Corresponding authors.
E-mail addresses: schaerer@smu.edu.sg (M. Schaerer), cduplessis@smu.edu.sg (C. du Plessis).
[1] Shared first authorship.
[2] See Appendix for a list of members of the Gender Audits Forecasting Collaboration.

https://doi.org/10.1016/j.obhdp.2023.104280
Received 10 September 2022; Received in revised form 18 August 2023; Accepted 4 September 2023
Available online 10 November 2023

(Brescoll, Uhlmann, & Newman, 2013), explicit (Glick & Fiske, 2001), implicit (Greenwald & Banaji, 1995), and normative factors (Fernandez-Mateo & Fernandez, 2016; Larwood, Szwajkowski, & Rose, 1988; Paluck, 2008; Vial, Brescoll, & Dovidio, 2019), all of which may remain stable or fluctuate across time. Research and theory on behavioral ecology and cultural transmission provides theoretical reasons to anticipate both stability and plasticity in gender discrimination across generations (Charlesworth & Banaji, 2022; Sng, Neuberg, Varnum, & Kenrick, 2018; Varnum & Grossmann, 2016, 2017).

According to widely influential theories of gender roles and gender inequality, social stereotypes both reflect and buttress women's and men's traditional roles in families and communities (e.g., caregiver roles). These in turn stem in part from physical differences between women and men, in particular women's role in birthing and nursing children (Eagly & Wood, 2012; Eagly, Wood, & Diekman, 2000). In contemporary societies, the legitimacy behind women rather than men serving in caregiver roles is much diminished. Yet, traditional gender roles, and associated explicit and implicit beliefs contributing to discrimination, can still persist (Banaji & Greenwald, 2013; Glick & Fiske, 2001). The transmission of cultural values across generations, even after the historical circumstances that gave rise to them have largely faded (Nisbett & Cohen, 1996; Talhelm et al., 2014), suggests that gender biases will perpetuate themselves across time (Cortes & Pan, 2018; Levanon & Grusky, 2016).

At the same time, there are both theoretical and empirical reasons to expect discrimination against women in particular to be less prevalent in the present than in the past (Eagly, Nater, Miller, Kaufmann, & Sczesny, 2020). Hierarchy-attenuating ideologies with increasing social legitimacy (e.g., feminism, egalitarianism; Alexander & Welzel, 2011; Inglehart & Welzel, 2005; Norris & Inglehart, 2004; Sidanius & Pratto, 1999; Welzel, 2014) may coexist with and motivate the correction of sexist biases (Crandall & Eshleman, 2003; Fazio, 1990; Gaertner & Dovidio, 1986). At an institutional and individual level, diversity and inclusion motives directed at women (Block, Croft, De Souza, & Schmader, 2019) represent a countervailing force that could overcome the influence of group stereotypes on selection decisions (Leslie, Manchester, & Dahm, 2017; Naumovska, Wernicke, & Zajac, 2020). In recent years, the #MeToo movement has shifted public norms related to gender (Johnson & Hawbaker, 2018; Kunst, Bailey, Prendergast, & Gundersen, 2019; Luo & Zhang, 2021; Soklaridis et al., 2018), eliciting growing sympathy for working women and perhaps by extension support for female job candidates.

Empirical investigations using an array of survey, indirect, experimental, and field methods have reached varying conclusions, including selective weakening of gender stereotypes and norms across time (Eagly et al., 2020; Goldin, 2006; Hammond, Milojev, Huang, & Sibley, 2018), stability or even increases in stereotypes (Haines, Deaux, & Lofaro, 2016), the subtle persistence of gender discrimination (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Rudman & Glick, 2001), advances in achieving equitable treatment (Ceci, Ginther, Kahn, & Williams, 2014; Landy, 2008), a pattern of decelerating cultural change (Bar-Haim, Chauvel, Gornick, & Hartung, 2018; Norris & Inglehart, 2004), and "reverse" gender discrimination against men (Card, DellaVigna, Funk, & Iriberri, 2021, 2023; Ceci, Kahn, & Williams, 2023; Williams & Ceci, 2015). Although likely due in part to the variety of methods employed and outcomes examined, these diverging conclusions have fueled a politically charged debate about the extent to which gender discrimination has persisted into the present (Arkes & Tetlock, 2004; Banaji & Greenwald, 2013; Banaji, Nosek, & Greenwald, 2004; Ceci et al., 2014; Ceci et al., 2023; Heilman & Eagly, 2008; Landy, 2008; Moss-Racusin et al., 2012; Williams & Ceci, 2015).

The present investigation seeks to introduce new evidence to this discussion by conducting a preregistered meta-analysis of 44 years of field audits of gender bias in callback rates for job applications (Study 1), and an accompanying forecasting survey gauging academic and lay predictions about the likely results (Study 2). Audit studies, in which job applications from carefully-matched female and male candidates are sent to real organizations, have high ecological validity and can estimate a causal effect of gender on hiring and selection decisions (Neumark, 2018; Quillian & Midtbøen, 2021; Quillian, Pager, Hexel, & Midtbøen, 2017; Rich, 2014). In contrast, observational field investigations, for example of performance evaluations, wage gaps, or job promotions, may be confounded by unmeasured differences between women and men (Card, DellaVigna, Funk, & Iriberri, 2020). Another option for studying gender bias are laboratory experiments that typically use hypothetical hiring scenarios and non-expert participants who may exhibit biased judgments that would not emerge among experienced and accountable decision makers (Tetlock & Mitchell, 2009). Contrarily, since they are aware of being studied, laboratory participants may correct their judgments for social desirability reasons (i.e., to avoid appearing prejudiced or sexist), shrouding biases that might have been observed under more naturalistic conditions (Tierney et al., 2020). In a meta-analysis of audits of real organizations that did not know they were part of a scientific study, the effects of year can be used to assess stability or change in labor market biases over time (Eagly, Makhijani, & Klonsky, 1992; Koch, D'Mello, & Sackett, 2015; Quillian & Lee, 2023; Quillian et al., 2017; Stanley & Jarrell, 1998; Williams & Tiedens, 2016). Our specific focus on field audits of the effects of candidate gender on selection decisions therefore maximizes ecological validity and causal inferences, both of which are critical for highly informative tests of cultural changes in discriminatory treatment.

2. Competing theories of stability and change in gender discrimination

One of our key research questions was whether there is a time trend in gender discrimination in job application outcomes. Different patterns of cultural evolution in discrimination based on applicant gender are possible. Biased selection decisions may have remained stable over time, such that there is significant discrimination against women in recent as well as older field audits. This *persistence-of-bias account* posits that the continuing existence of many stereotypes and sexist beliefs (Eagly et al., 2000; Glick & Fiske, 2001) means behavioral discrimination should continue largely undiminished. Indirectly relevant meta-analytic evidence suggests that in many Western societies, racial and ethnic discrimination in selection for jobs has persisted across all observed time periods (Quillian & Lee, 2023; Quillian et al., 2017). If racial discrimination in hiring remains pervasive, this increases the plausibility that gender bias in candidate selection, which theoretically derives from some of the same implicit and explicit mental processes (Greenwald & Banaji, 1995) and situational forces (Larwood, Szwajkowski, & Rose, 1988), remains widespread as well. More direct evidence is provided by recent work demonstrating that gender stereotypes remain deeply ingrained in the minds of many in the form of automatic associations (Charlesworth & Banaji, 2022) and are reflected in widely consumed cultural products such as music (Boghrati & Berger, 2023). If gender stereotypes are "in the air" in the surrounding culture and conditioned in people's minds, it is reasonable to expect that stereotype-based discrimination in selection decisions against female and male candidates is commonplace as well. Substantial preceding research thus provides a strong *a priori* empirical reason to expect similar robust biases in hiring against female applicants, all the way up to the present.

Alternatively, discrimination against female candidates may have faded away over time, such that recent studies will reveal little gender disparity in selection. This *fading-of-bias account* acknowledges that unfair discrimination was common in past generations, contributing to inequalities that have carried over into the present. For example, gender gaps in representation in senior leadership positions today are attributable in part to upstream biases in selection decades ago that limited the present-day pool of available talent just below the executive level. Yet from this perspective, today's organizational decision makers have become better at correcting for societal stereotypes when it comes to

deciding who to hire (Tetlock & Mitchell, 2009), and given empirical evidence of changes in at least some gender norms and behaviors (Badura, Grijalva, Newman, Yan, & Jeon, 2018; Hora, Badura, Lemoine, & Grijalva, 2021; Koenig, Eagly, Mitchell, & Ristikari, 2011) may also be less biased in the first place. A more nuanced view posits that gender discrimination is uncommon in hiring decisions, which are publicly visible and carefully monitored for bias by individuals and organizations, but still visible in compensation decisions that occur behind a shroud of confidentiality (Ceci et al., 2023). Regardless, from this perspective, contemporary selection processes are in the aggregate no longer substantially impacted by applicant gender.

Yet another possibility is that gender preferences in hiring decisions have reversed over time. This would imply a negative time trend for discrimination against women initially, followed by a transition into a preference for female candidates in recent years as organizations have striven to overcome historical discrimination and contemporary underrepresentation. Under the "*reverse" discrimination* account, some individuals and organizations perceive female employees as offering diversity value that goes above-and-beyond their human capital value (e.g., Chang, Milkman, Chugh, & Akinola, 2019; Leslie et al., 2017). Whether a matter of genuine inclusion motives or strategic signaling, there could be a premium associated with female candidates for certain roles in contemporary organizations. This may be especially true for roles from which women were historically excluded and where they continue to be underrepresented. In such contexts, the motive to include more female candidates and achieve greater representation of women should be stronger. Note that the average gender discrimination effect contextualizes any time trend. For a society to collectively exhibit bias against women for most of its history and then show progressively more gender-balanced judgments over time is quite different from initially gender-neutral judgments turning into a preference for female candidates. The former is a case of the gradual fading-of-bias; the latter, a gradual introduction of a different bias.

Finally, it is possible that the trend across time will reveal an inflection point associated with recent social movements related to workplace sexual harassment, specifically the #MeToo movement (Luo & Zhang, 2021). The global attention to prominent harassment and assault cases, along with the powerful everyday narratives associated with #MeToo shared on social media, may have accelerated cultural change processes with regard to gender. This *#MeToo hypothesis* expects more favorable outcomes for female applicants in the post #MeToo years (i.e., from 2018 onwards; see Luo & Zhang, 2021). Although the direct focus of #MeToo is on gender-based harassment, the accompanying changes in gender sensitivities and standards for appropriate behavior may have spilled over to other forms of gender discrimination, such as in selection decisions.

In testing for potential cultural changes, we consider the gender typicality of the job (stereotypically female, relatively gender balanced, or stereotypically male), since past laboratory and field studies identify this as a key moderator of hiring evaluations (Davison & Burke, 2000; Eagly et al., 1992; Glick, Zion, & Nelson, 1988; Koch et al., 2015; Riach & Rich, 2002). The theoretically predicted patterns regarding the potential persistence, fading, and reversal of bias in selection decisions do not necessarily apply to jobs that society has historically deemed the purview of women (e.g., nurse or receptionist) as contrasted with male-typed (e.g., construction worker or carpenter) and comparatively gender-balanced (e.g., sales representative) jobs. At the same time, selection decisions for female-typed jobs are of theoretical interest because they could reflect a general weakening of social stereotypes and gender norms, if employers are increasingly open to men who apply to fulfill such roles in organizations. Conversely, a reduction in discrimination against female candidates for male-typed and gender-balanced jobs without a simultaneous increase in selecting men for female-typed jobs would more likely reflect selective changes in stereotypes (Eagly et al., 2020) and employers seeking to increase the representation of women but not men (Block et al., 2019).

This led to a set of research questions for which the meta-analysis aimed to help adjudicate between the competing theories. Similar to Tierney et al. (2020, 2021), who engaged in competitive theory testing in the context of gender bias, we carried out a single set of pre-registered analyses whose results could support or fail to support different theoretical accounts with contrasting hypotheses.

Research Question 1: On average, do men experience more positive job application outcomes than women?

Research Question 2: Is the effect of gender on job application outcomes moderated by the job's gender stereotypicality?

Research Question 3: Is there a time trend in gender bias in selection decisions?

Research Question 4: Are the years since the onset of the #MeToo movement (i.e., from 2018 onwards) associated with a change in discrimination against female applicants?

Although it is admittedly speculative and may or may not find empirical support, the #MeToo hypothesis can be tested with the available data and is theoretically informative regarding the nature of changes in gender norms (linear or nonlinear, consistent or fragmented). More generally, empirical investigations suggest interrelationships and spillovers between only indirectly related dimensions of cultural change (Charlesworth & Banaji, 2022; Charlesworth, Yang, Mann, Kurdi, & Banaji, 2021; Norris & Inglehart, 2004; Varnum & Grossmann, 2016, 2017). This provides at least some prior empirical and theoretical basis to expect that shifts in societal norms regarding sexual harassment could spill over to selection decisions involving female and male candidates.

Finally, as an exploratory analysis, we examined whether there was either a preference for male applicants, a preference for female applicants, or no bias toward either gender in recent years. The average effect speaks to whether organizations should focus their debiasing interventions on the selection process or further downstream such as in compensation decisions, work assignments, and promotions. It also speaks to whether societies seeking greater gender balance in the workplace should invest their energies in preventing selection-stage bias by employers or focus further upstream on access to educational opportunities and childcare. However, how precisely to parse the last half-century into different time periods on a meaningful basis is not immediately clear. For example, one could divide studies by decade, into 5-year spans, pre-and-post 2000, or into quartiles based on the total number of investigations. Thus, although the theoretically predicted patterns of overall discrimination past and present were pre-registered, our statistical analyses regarding the presence or absence of bias in recent time periods were based on arbitrary time increments and are thus exploratory in nature.

The present investigation's contributions to the literature on gender are multifold. We assess the direction, severity, and stability of gender discrimination with unprecedented rigor, leveraging recent open-science best practices such as pre-registration of methodology and analyses (Wagenmakers et al., 2012) and an audit by a "red team" of external experts (Lakens, 2020) to prevent researcher bias. We pre-specify the competing empirical predictions of sometimes complementary, and sometimes contradictory theoretical accounts, maximizing the informational value of the investigation for theories of gender and society. Our substantial sample of 44 years of audit studies on gender, the largest ever assembled, allows for informative tests of not only the moderating role of job stereotypicality but also recent events such as the #MeToo movement. This is the first investigation to assess scientist and lay perceptions of gender discrimination over the years and map these on to objective empirical results, offering the opportunity to put clashing priors about societal change and pervasive prejudice to a rigorous empirical test (Tetlock, Mellers, Rohrbaugh, & Chen, 2014). The Trajectory of Discrimination project is part of a broader, ongoing program of research from our group that seeks to open the science of diversity and discrimination using recent open and crowd science innovations (Dreber et al., 2015; Klein et al., 2014; Lakens, 2020; Wagenmakers et al., 2012).

3. Study 1: A meta-analysis of stability and change in gender discrimination over time

Our empirical approach for the meta-analysis followed a multi-step strategy. Before committing ourselves to our methodology, we recruited a "red team" (Lakens, 2020) of expert critics to provide detailed feedback to the main project team (blue team) regarding the initial project plan. The revised and optimized approach was then pre-registered on the Open Science Framework (https://osf.io/ha3n4). Building on meta-analytic investigations that focused on recent studies only (e.g., Lippens, Vermeiren, & Baert, 2021), or that sampled mostly laboratory experiments along with a smaller set of field audits (Koch et al., 2015), we attempted to identify all field audits from any year concerned with gender and hiring discrimination. Next, we utilized an *a priori* coding scheme to extract and process relevant information from the target articles and reports to create a database for our analyses. Finally, we conducted the preregistered meta-analytic analyses, as well as additional exploratory analyses.

3.1. Methods

3.1.1. Red team approach

The prevalence of gender bias in hiring and other forms of group-based discrimination are among the most controversial issues in the social sciences (Arkes & Tetlock, 2004; Banaji et al., 2004; Ceci et al., 2014; Ceci et al., 2023; Heilman & Eagly, 2008; Landy, 2008). Concerns about potential researcher ideological and intellectual commitment biases on both sides are common in this space (Clark & Winegard, 2020; Cyrus-Lai et al., 2022; Duarte et al., 2015; Jost et al., 2009). In light of the strong need to enhance objectivity, increase trust, and generally maximize the informational value of our meta-analysis, we leveraged emerging best practices of open science, including pre-registration of analyses and open data (Nelson et al., 2018; Wagenmakers et al., 2012). This constrains researcher degrees of freedom, and greatly expands opportunities for re-analyses and alternative perspectives from other scholars.

To further optimize our methods, we employed the innovative new "red team" approach (Lakens, 2020; Zenko, 2015). A red team is a designated team of scientific experts external to the core author group (the "blue team"). Two coordinators recruited an independent team of experts on statistics, meta-analysis, and gender research, as well as a librarian, to critique all aspects of our meta-analysis plan, point out potential issues, and suggest improvements. The goal of the red team approach was to improve the quality of the research project by identifying flaws and challenging dominant assumptions in our work, incorporate different viewpoints, and invite early feedback from international experts. We preregistered and carried out the optimized study methodology and analysis plan, followed by another round of feedback from the red team.

Unlike traditional peer reviewers, red team members are financially compensated for their work and provide feedback throughout the project, when it is still possible to correct errors or methodological weaknesses. The logic of the red team approach is comparable to a registered report publication system, in which research protocols are reviewed by the journal before the results are known (Chambers et al., 2015). However, the criticism is not invited by the journal but by the authors (blue team). Such an approach allows for an exchange between researchers and a "devil's advocate" that aims to produce a higher quality research plan before submission to a journal by identifying oversights, soliciting feedback from experts, and preventing groupthink (Lakens, 2020). A red team is also similar in some respects to an adversarial collaboration, where researchers with directly opposing predictions work to design a study together (Clark & Tetlock, 2022; Mellers, Hertwig, & Kahneman, 2001), except that red team members are recruited for expertise alone rather than their intellectual committments.

Because our goal was to generate critical feedback on our bibliographic search, data coding, analysis, as well as our theorizing and inferences, we recruited five red team members (four female, one male) with expertise in one or more of these domains (see Supplementary Online Materials for anonymized brief profiles). Three of the red team members were scholars with training and publishing experience in the domain of gender research, some with additional expertise in field audit methods, and included one qualitative gender studies expert. This was complemented by a scholar with expertise in meta-analytic methods and statistics, as well as a senior librarian who advised us on our bibliographic search approach. With the exception of one scholar who was in advanced doctoral training and the librarian, the remaining red team members had doctoral degrees in their respective areas and were either post-doctoral fellows or tenure-track faculty. Four red team members received financial compensation for their feedback and one red team member declined payment.

We solicited feedback from the red team at two stages of the project. An initial round of feedback was requested after we had conducted a preliminary bibliographic search, developed a preliminary coding scheme, and extracted data from a portion of the studies. Five red team members participated in the first stage. No analyses had been conducted at this time. A second round of feedback was requested after completion of the revised search, data analyses, and draft manuscript. Three red team members participated in the second round (one gender expert, one statistician, and one librarian). In both rounds, the red team was given approximately two to four weeks to provide the blue team with written feedback. After receiving the first round of feedback on the planned methods, we made extensive revisions to our approach and responded to each suggestion by the red team, explaining what changes were made to address their concerns or why we decided not to incorporate a particular suggestion. For example, based on the first round of feedback of the red team, we revised our search terms and preregistered our revised search and coding approach in detail. After receiving the second round of feedback, we made changes to the manuscript (e.g., clarify arguments, extend discussion) and Supplementary Online Materials (e.g., provide more methodological detail, conduct additional analyses). For instance, we conducted additional publication bias analyses and added more material on potential limitations of the current set of studies. The red team also identified a coding error and a rounding error which we subsequently corrected. The full-length, anonymized feedback by the red team and the blue team's respective responses are available on the Open Science Framework (https://osf.io/pt4gn). Table S3 in the Supplementary Online Materials provides an overview of the most important feedback exchanges for each aspect of the meta-analysis.

3.1.2. Identification of relevant studies

Once the blue team and red team had settled on the meta-analysis methodology and planned statistical analyses, we worked to identify all published and unpublished field audits examining a contrast in hiring-related outcomes between female and male job applicants. This includes all in-person audit studies and resume correspondence studies that manipulated gender either "within" employer (i.e., an employer received applications from both female and male candidates) or "between" employer (i.e., an employer received applications from either a female or male candidate) and that kept all other candidate characteristics equivalent either through randomization or creating matched pairs.

We employed multiple search strategies during April 2021, including searches in academic databases, citation searches, email requests to corresponding authors of gender-related field experiments, and public calls for unpublished work. First, we conducted a systematic search of primary academic databases, including Web of Science Core Collection (A&HCI, BKCI-SSH, BKCI-S, ESCI, SCI-EXPANDED, SSCI), Business Source Ultimate (via EBSCO), EconLit (via EBSCO), Humanities International Complete (via EBSCO), APA PsycArticles (via EBSCO), APA PsycInfo (via EBSCO), SocINDEX (via EBSCO), and Google Scholar (first

1,000 results). Our search string, expanded substantially after feedback from the red team, consisted of a combination of keywords related to gender (e.g., *gender, sex*, female**), discrimination (e.g., *bias, stereotyp*, discriminat**), and field experimental methodology (e.g., *audit stud*, field experiment*, randomized trial**) with some variation depending on the search functions of the respective database. See the meta-analysis preregistration on the Open Science Framework (https://osf.io/ha3n4) for the exact search strings for each database and Table S1 in the Supplementary Online Materials for deviations from the preregistered protocol.

Second, we conducted backward and forward citation tracing to identify additional studies. Specifically, we reviewed the references of a number of important published articles, reviews, and meta-analyses related to gender discrimination (Adams, Gupta, & Leeth, 2009; Baert, 2018; Bertrand & Mullainathan, 2004; Bowen, Swim, & Jacobs, 2000; Hebl, Cheng, & Ng, 2020; Jones, Peddie, Gilrane, King, & Gray, 2016; Koch et al., 2015; Moss-Racusin et al., 2012; Roth, Purvis, & Bobko, 2012; Tosi & Einbender, 1985; Triana, Gu, Chapa, Richard, & Colella, 2021) and reviewed the Google Scholar citations of the five most highly cited articles (Ahmed, Andersson, & Hammarstedt, 2013; Correll & Benard, 2007; Gaddis, 2015; Neumark, Bank, & Van Nort, 1996; Rivera & Tilcsik, 2016) of the studies identified through our academic database search.

Third, we took additional steps to identify unpublished and "in press" studies. We searched for unpublished dissertations related to our topic of interest on ProQuest Dissertations and Theses Global. We also issued public calls via listservs, discussion forums, and social media pages of relevant academic communities (e.g., Academy of Management Organizational Behavior Division and the Gender and Diversity in Organizations Division, American Sociological Association, PsychMap, PsychMethods). Finally, we contacted the corresponding authors of studies identified via the systematic search of academic databases and citation tracing described above to directly request information about any unpublished studies.

Our search produced a total of 6,754 search results. Using a bibliographic management software (Zotero), we excluded 709 duplicate articles and three retracted articles. One blue team author subsequently assessed each of the remaining 6,042 results for relevance ("yes", "no", "maybe") based on title and abstract using a web-based, collaborating screening platform (Rayyan) which helps organize and manage collaborative systematic literature reviews. Those coded as "maybe" were assessed by a second author. For the resulting 456 records, we subsequently retrieved the full-text articles for more careful examination. Twelve articles could not be retrieved, leaving 444 articles for full-text examination. Following our preregistered inclusion criteria, we excluded additional articles because they did not contain field experimental data ($n = 193$), gender was not investigated or properly randomized ($n = 99$), no hiring outcomes were measured or reported ($n = 30$), relevant statistics were missing and could not be provided by the authors ($n = 19$), or the underlying data were the same as in another article ($n = 18$). The final sample included 85 usable field studies and is thus more comprehensive and up-to-date than earlier meta-analyses on gender discrimination (Koch et al., 2015; Lippens et al., 2021), although of course also building on this important prior work. Fig. 1 contains the PRISMA flow diagram (Page et al., 2021) which summarizes the overall search process. Fig. 2 depicts the number of audit studies across geographic regions and time.

3.1.3. Data extraction

We coded key characteristics of each study according to a preregistered coding rubric (see Open Science Framework: https://osf.io/ha3n4). As the coding progressed, we further refined the coding scheme where necessary. Deviations from the preregistered protocol are reported in Table S1 (see Supplementary Online Materials). For example, rather than coding whether a study used a matched pairs design or not, we decided that it was more meaningful to separately code a) whether the study manipulated gender within or between employers and b) whether the female and male applications were real, manually matched pairs (e.g., two trained actors or two real resumes of similar quality) or equivalent, fictitious pairs (e.g., the same resume manipulated to have either a female or male candidate name). The final coding rubric is reported in Table S2.

The coding involved information at both the study and effect level. Study level characteristics are constant for the entire study, such as gender ratio of the author team or year of data collection. For some studies, multiple characteristics were coded at the effects level. For example, a study that separately reported callback data across three countries would produce three effect sizes, and a study that separately reported callback data for eight professional groups (e.g., cleaner, clerk, gardener) would result in eight effect sizes.

Following the preregistered protocol, all objective variables (e.g., data collection year, applications sent, callbacks) were coded by one author and subsequently verified for accuracy by a research assistant. In case of disagreement, further investigations were conducted to verify that the extracted information was accurate. Another author was consulted to resolve ambiguities. For the gender variable, we followed the established approach of other meta-analyses on gender and ethnicity to extract the main effect comparing overall discrimination between cisgender[3] female and male applicants (e.g., Flage, 2018; Lippens et al., 2023; Koch et al., 2015; Quillian et al., 2017; Quillian & Lee, 2023; Zschirnt, & Ruedin, 2016). For example, if a study orthogonally manipulated gender (female vs. male) and age (younger vs. older; e.g., see Baert et al., 2016), we examined gender differences across both younger and older applicants combined, as these characteristics naturally vary in labor markets. For our time variable, we extracted the year in which the data were collected. If data collection spanned multiple years, we extracted the year in which most of the applications were sent out.

For our subjective variable, gender typicality of job according to broader cultural stereotypes, we used a preregistered approach employing human coders for the main analysis. We complimented this with an exploratory approach based on objective country-level demographic data on gender representation in particular jobs. The human coder approach allows for a deeper and more consistent level of granularity, as country-level data may not be consistent and equally granular across countries. The objective country-level data may remove potential coder bias and more accurately capture cross-national differences in gender representation for the same type of job and account for within-job shifts over time. For the preregistered human coder approach (Derous & Ryan, 2012), four authors independently coded studies (intercoder agreement was substantial, Fleiss kappa = 0.77, $p < .001$; see Table 1 for example jobs for each gender category) and discrepancies were resolved through a majority vote approach or discussion (if there was no majority). The coders who categorized jobs based on their stereotypicality were from the following nations: the United States and Chile, Switzerland, Vietnam, and Australia and South Africa. For the objective data approach, we retrieved country-level gender representation data via each country's official labor statistics reports (if available), the United Nations website, or other governmental/non-profit reports. Following past research (e.g., Hora et al., 2021; Koch et al., 2015), a job was coded as female-typed (male-typed) if the representation of women (men) was 65% or higher, and coded as gender-balanced otherwise (see Table S5 in the Supplementary Online Materials for sensitivity analyses

[3] In our investigation, we focused on the comparison between cisgender females and males as this has been the primary comparison in past research to date. One study by Granberg et al. (2020) also included transgender conditions in addition to the female and male cisgender conditions. For the present research, we focused on the latter two conditions as there were not enough studies systematically examining transgender candidates at a meta-analytic level. We encourage future research to conduct more systematic investigations on this important topic.

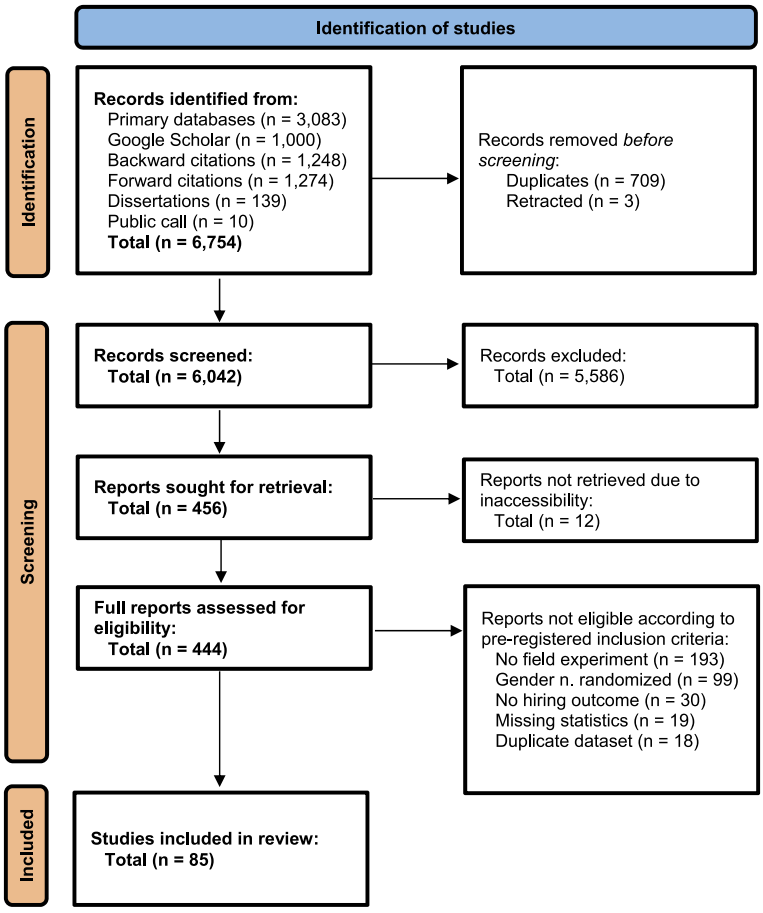


Fig. 1. PRISMA flow diagram. The diagram depicts the information flow through the different phases of our systematic review, including the number of records identified, included and excluded, and the exclusion reasons.

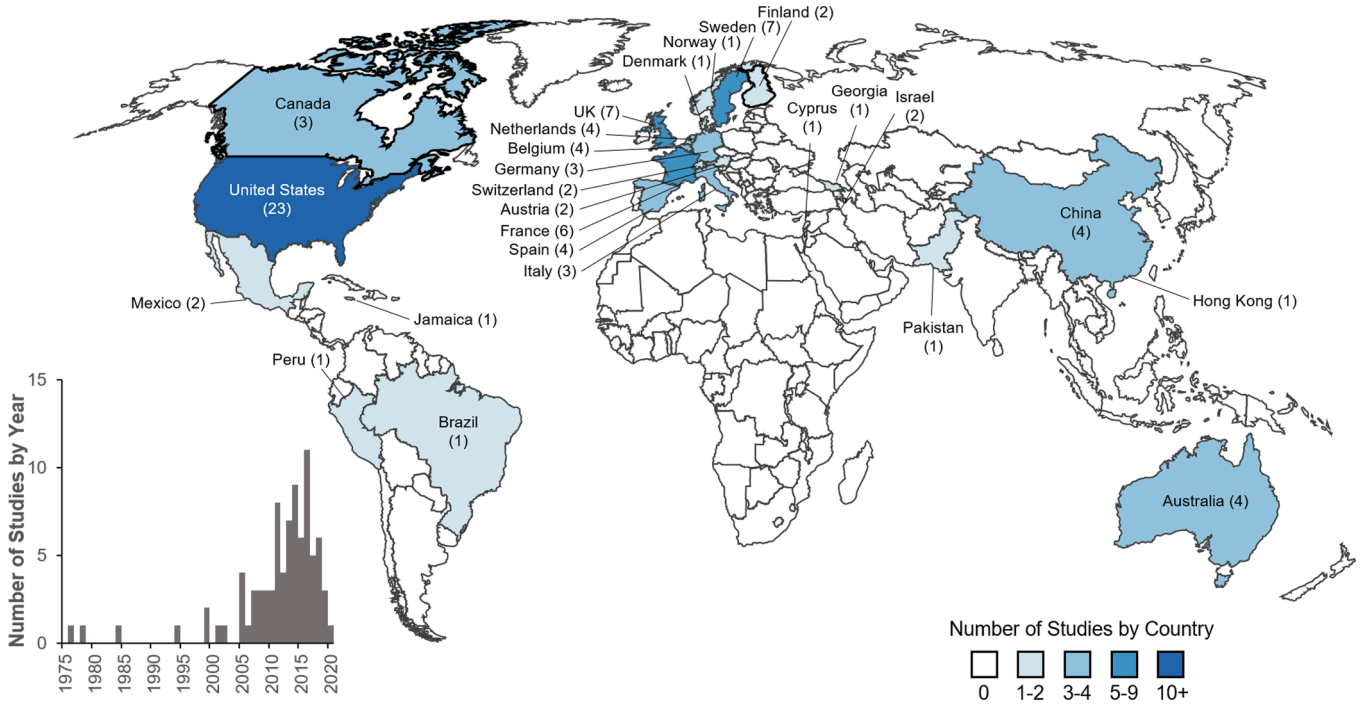


Fig. 2. Number of audit studies across geographic regions and across time. World map visualizing the number of field audits included in our sample across different countries and territories (center) and across year of data collection (bottom left).

Table 1
Example Jobs by Gender Typicality. The table provides example jobs for each category of job gender typicality (female-typed, gender-balanced, male-typed) as categorized by four human raters. Example jobs are presented in alphabetical order within each category.

Female-Typed Jobs	Gender-Balanced Jobs	Male-Typed Jobs
Administrative clerk	Accountant	Auto mechanic
Executive secretary	Baker	Carpenter
Hairdresser	Barkeeper	Chef
HR professional	Call center worker	Computer specialist
Nurse	Commerce worker	Construction worker
Payroll clerk	Graphic designer	Electrician
Primary school teacher	Laboratory worker	Engineer
Receptionist	Marketing technician	Gardener
Social worker	Sales representative	Plumber
Waitstaff	Secondary school teacher	Truck driver

using alternative cutoffs of 60% and 70%). The association between the subjective and objective coding approaches for gender typicality of jobs was strong (Cramer's V = 0.71, p <.001).

To assess country-level gender inequality, an additional variable (Gender Inequality Index, or GII) was retrieved from the United Nations Human Development Report (United Nations Development Programme, 2020). The GII is a composite measure of gender inequality using data on reproductive health (e.g., maternal mortality), empowerment (e.g., women with higher education degrees), and the labor market (participation of women in the labor force). A low (high) GII value indicates low (high) inequality between women and men. The GII was published every five years between 1995 and 2010 and annually between 2010 and 2019. Thus, for studies published before 2010, we took the GII index with the smallest temporal distance to the data collection year (e.g., for a study from 1999, we took the GII from the 2000 report).

In cases of missing data, we followed the preregistered protocol and reached out to the corresponding author of the respective study. In most cases, the authors were able to provide us with the missing data (e.g., year of data collection, callback rates). When missing data could not be obtained from the authors, the study was either excluded (e.g., when we could not compute an effect size for the study) or we made reasonable assumptions (i.e., for one study, we inferred callback rates from figures).

3.1.4. Statistical analyses

Data from each study were a 2x2 frequency table as shown below.

		Outcome	
		Success	Failure
Applicant Gender	Female	A	B
	Male	C	D

The effect size measure of interest in the meta-analyses was the log odds ratio. Before computing the log odds ratios, we added 0.5 to all cells of the 2x2 frequency table to decrease bias in the estimator of the log odds ratio and to avoid division by zero when computing the log odds ratio and its sampling variance in cases where some of the cells equaled zero (Walter & Cook, 1991). The log odds ratio of each study was computed using (equations 11.57 and 11.58 in Borenstein & Hedges, 2019),

$$\ln\left(\frac{AD}{BC}\right)$$

where ln denotes the natural logarithm. The corresponding sampling variance of the log odds ratio was computed using (equation 11.59 in Borenstein & Hedges, 2019),

$$\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}$$

We preregistered to conduct univariate random-effects meta-analyses for testing each hypothesis. During data collection, we realized that many studies contributed more than one effect size based on an independent sample to the meta-analysis; the minimum, maximum, median, and average number of effect sizes a study contributed was 1, 42, 1, and 2.9, respectively. Hence, we decided to deviate from the preregistered analyses and take the nesting of effect sizes within studies into account by conducting three-level multilevel meta-analyses (Konstantopoulos, 2011; Van Den Noortgate & Onghena, 2003). A three-level multilevel meta-analysis adds an extra level to the meta-analysis model to take the nesting structure into account. We report the results of the multilevel meta-analyses in the main report, but also present the results of the preregistered univariate random-effects meta-analyses in the Supplementary Materials. Parameter estimation in the multilevel meta-analyses was done using restricted maximum likelihood estimation. For each multilevel *meta*-analysis, the I^2-statistic (i.e., the proportion of total variance that can be attributed to heterogeneity, Higgins & Thompson, 2002) was computed analogous to what is described in Nakagawa and Santos (2012). We evaluate whether the normality assumptions of the multilevel meta-analysis model hold in the Supplementary Online Materials (see Section S7 and Fig. S1).

All analyses were conducted in the statistical software R (R Core Team, 2021, version 4.1.1). The R package "metafor" (Viechtbauer, 2010, version 3.0.2) was used for conducting the multilevel meta-analyses and some of the data visualizations and the R packages "weightr" (Coburn & Vevea, 2016, version 2.0.2) and "puniform" (van Aert, 2021, version 0.2.4) were used for publication bias analyses. Two-tailed hypothesis tests were conducted using $\alpha = 0.05$ and 95% confidence intervals were computed. R code and analysis output is available on the Open Science Framework (https://osf.io/pt4gn/).

To account for potential confounding factors, we included several control variables in our analyses. First, to account for the possibility that more recent studies have been conducted in regions with more egalitarian gender norms, we controlled for national scores on the GII. Second, it is possible that the experimental designs of audit studies have become more complex, such as the increased use of multifactorial designs manipulating gender and an intervention designed to reduce discrimination (Byrd, 2019). Collapsing across baseline and intervention conditions could depress discrimination effect sizes over time. To take this into account, we controlled for whether studies manipulated a single or multiple experimental factors. Third, we accounted for the gender composition of the author teams over time. Note that we preregistered to also include application method as a control variable, but submitting a resume was used for nearly all effect sizes as the application method (98.4%). Due to a lack of variance, we did not include this variable as a control variable in the meta-analyses.

The results of the multilevel meta-analyses and univariate random-effects meta-analyses are reported in Table 2 (multilevel meta-analyses) and S4 (univariate random-effects meta-analyses; see Supplementary Online Materials), respectively. Research Question 1 regarding overall preference for female vs. male candidates across the years was tested by conducting a meta-analysis without any predictor variables as an intercept-only model (see Models 1a/b). Research Question 2 examining moderation by job type was tested by including a dummy variable for non-female-typed jobs (a score of 0 implies that a job was a stereotypically female-typed job, and a score of 1 indicates either a stereotypically male-typed job, or a gender-balanced job or set of jobs; see Models 2a-d). For our main analyses, we followed the preregistered protocol to collapse male-typed and gender-balanced typed jobs into a single category to maximize the statistical power of tests of discrimination against women across time. Grouping gender-balanced occupations (e.g., accountant) with male-typed rather than female-typed jobs was further supported by research on gender as a diffuse status cue (Ridgeway, 1991), traditional cultural stereotypes of women as less competent than men (Fiske, Cuddy, Glick, & Xu, 2002), and pervasive implicit and linguistic stereotypes regarding career versus family roles (Charlesworth & Banaji, 2022; Charlesworth et al., 2021), all of which predict biases in favor of men and against women even for

comparatively neutral professional settings and work tasks (Jost, 1997; Pelham & Hetts, 2001). However, we also present results separately by male-typed, gender-balanced, and female-typed jobs. The competing theoretical predictions regarding Research Question 3 (persistence vs. fading-of-bias) were tested by including the year of application in the meta-analysis. This variable was first centered by subtracting the year of the oldest application (i.e., 1976) to avoid convergence issues. The centered variable was included as predictor in the meta-analyses (see Models 3a-d). Research Question 4 regarding #MeToo was tested by comparing the years 2018–2021 to 2014–2016,[4] and our analyses regarding the extent of gender bias in recent versus distant time periods were carried out by breaking the studies into different yearly intervals on an exploratory basis (see Fig. 5).

3.2. Results

3.2.1. Is there evidence of gender discrimination favoring male or female applicants?

We first examined whether female or male applicants experience more positive job application outcomes overall (i.e., across all job types and years). Across all 85 studies from 1976 to 2020, the average odds of male applicants to receive a callback was 0.91 times the odds of equally qualified female applicants (95% confidence interval ranged from 0.86 to 0.97, z = -3.00, p =.003; see Table 2, Model 1a). Importantly, heterogeneity of the true effects was large (I^2-statistic = 82.8%), implying that 82.8% of the total variance can be attributed to heterogeneity. This is also apparent in the wide 95% prediction interval (0.49 to 1.70), which indicates the effect size for a future study from the same distribution of true effects. The heterogeneity could not be explained by including the control variables (inequality index, presence of moderators in study, proportion of female authors), because the heterogeneity of the true effects remained large (I^2-statistic = 83%). However, the effect of candidate gender was no longer significant (z = -0.15, p =.883; see Table 2, Model 1b) when we included the control variables.

3.2.2. Does gender discrimination depend on the gender-typicality of the job?

To further examine whether job application outcomes for female and male applicants varied as a function of the gender typicality of the job category that was applied for, we tested whether job application outcomes are less favorable for women for the combined categories of male-typed and gender-balanced jobs, relative to female-typed jobs. Using the human coded gender-typicality variable, job type significantly moderated the effect of gender on callbacks (0.26, z = 4.91, p <.001; see Table 2, Model 2a). Specifically, the average odds of a male (vs. female) applicant to receive a callback was significantly lower for female-typed jobs (odds ratio: 0.75, 95% confidence interval: 0.68, 0.83) compared to male-typed and gender-balanced jobs combined (odds ratio: 0.97, 95% confidence interval: 0.91, 1.03). The moderator remained significant when we included the control variables (0.25, z = 4.85, p <.001; see Table 2, Model 2b). The results were consistent when we used the objective country-level data as gender-typicality moderator: job type significantly moderated the effect on gender on callbacks, excluding (2.68, z = 4.96, p <.001; see Table 2, Model 2c) and including (2.68, z = 4.89, p <.001; see Table 2, Model 2d) the control variables. Thus, female applicants were on average more likely than male applicants to receive callbacks for female-typed jobs, while there was no significant effect of candidate gender for male-typed and gender-balanced jobs overall. Including this variable still left a large amount of heterogeneity (residual I^2-statistics ranged from 80.8% to 81.3%).

3.2.3. Has gender discrimination changed over time?

One of our key research questions for the meta-analytic investigation was whether there has been stability or change in gender discrimination over time for male-typed and gender-balanced jobs considered together. To test this, we fitted a multilevel meta-regression model with the year in which the applications were sent out as a predictor. Using the preregistered human coded gender-typicality variable, there was a significant, albeit small, decreasing time trend of the average log odds ratio (-0.010, z = -2.56, p =.011, residual I^2-statistic = 81.2%; see Table 2, Model 3a), suggesting that job application outcomes for female candidates improved over time relative to male candidates. Including the control variables in the meta-regression model did not change the direction or significance of the time trend (-0.015, z = -3.01, p =.003, residual I^2-statistic = 80.7%; see Table 2, Model 3b). Fig. 3 shows that female applicants had a disadvantage over male applicants before 2009 and that this difference was no longer noticeable or, if anything, slightly reversed in direction starting in 2009. The trend also remained significant when we used the objective country-level data to identify non-female-typed jobs, both excluding (-010, z = -2.49, p =.013; see Table 2, Model 3c) and including (-0.013, z = -2.61, p =.009; see Table 2, Model 3d) the control variables. Jointly, these results demonstrate that the increasingly positive outcomes for female applicants over time are not likely attributable to the subjective vs. objective measurement of the gender typicality moderator, shifts in location, changes in study designs, or gender composition of research teams.

Although the time trend appears visually to be most pronounced for relatively gender-balanced jobs (Fig. 4), an exploratory analysis revealed that there was no significant interaction between time trend and job type (-0.01, z = -1.19, p =.236), meaning that we could not reject the hypothesis that the trend was the same across the three job types. In an exploratory analysis, we further broke down the outcomes for each job category by different time periods (Fig. 5A/B). Prior to 1991, we observed a preference for male applicants for male-typed and gender-balanced jobs, although these early intervals are based on a small number of studies and not significant. In more recent time periods (post 2009), we observed a preference for female candidates for gender-balanced jobs whose significance depended on the specific years in question, a significant preference for female applicants for female-typed jobs, and no significant gender-of-candidate preference for male-typed jobs.

In addition to the seemingly gradual shift over time (Fig. 3), an exploratory comparison of 2018–2020 relative to the preceding years 2014–2016 did not reveal a significant reduction in discrimination between those two time periods (z = 0.379, p =.704). This indicates that increasing support for female applicants is a longstanding trend and cannot be attributed to a sudden spike in support for female applicants due to the #MeToo movement that became a global phenomenon in 2017.

Overall, the fading-of-bias account's predicted decline in discrimination against women over time was supported (Research Question 3), as was moderation of gender discrimination by the stereotypicality of the job (Research Question 2). In contrast, the speculative possibility that the #MeToo years would be associated with accelerated cultural change (Research Question 4) did not find empirical support. Overall anti-female bias in selection decisions was not observed, and although some suggestion emerged of an aggregate anti-male bias this was not robust to covariates (Research Question 1). For our exploratory analyses regarding the presence of bias in recent years, the results are contingent on arbitrary and post hoc decisions regarding how intervals of years are divided and thus provide no robust evidence of contemporary gender discrimination for most jobs. Taken together, the results support the fading-of-bias account for male-typed and gender-balanced jobs (i.e., non-female-typed jobs), and the persistence-of-bias account for female-typed occupations.

[4] The year 2017 was excluded as the #MeToo trend started partway through 2017 (Luo & Zhang, 2021) which is therefore ambiguous as to which side of the cultural event it falls on.

Table 2

Results of hypothesis tests using multilevel meta-analyses. For each model and variable, the parameter estimate, standard error (), 95% confidence interval (CI) [], *z*-value, and corresponding two-tailed *p*-value of the multilevel meta-analyses are displayed. All models were fitted using restricted maximum likelihood estimation. $\widehat{\sigma}_1^2$ = the between-study variance in true effect size; $\widehat{\sigma}_2^2$ = the variance in true effect size of effect sizes nested in studies.

			Human Coding Approach (preregistered)				**Objective Data Approach** (65% cutoff; exploratory)			
	Average Discrimination		**Gender Typicality of Job**		**Time Trend**		**Gender Typicality of Job**		**Time Trend**	
	Model 1a	Model 1b	Model 2a	Model 2b	Model 3a	Model 3b	Model 2c	Model 2d	Model 3c	Model 3d
Intercept	-0.091 (0.030) [-0.151;-0.032] z = -3.000 ***p* =.003**	-0.017 (0.116) [-0.244;0.209] z = -0.148 ***p* =.883**	-0.283 (0.049) [-0.379;-0.188] z = -5.796 p <.001	-0.214 (0.119) [-0.448;0.020] z = -1.792 p =.073	0.325 (0.140) [0.051;0.600] z = 2.326 p =.020	0.356 (0.150) [0.061;0.650] z = 2.363 p =.018	-0.299 (0.052) [-0.401;-0.197] z = -5.728 p <.001	-0.221 (0.125) [-0.467;0.024] z = -1.766 p =.077	0.301 (0.137) [0.032;0.570] z = 2.196 p =.028	0.344 (0.151) [0.047;0.641] z = 2.271 p =.023
Non-female-typed job			0.255 (0.052) [0.154;0.357] z = 4.914 ***p* <.001**	0.255 (0.053) [0.152;0.358] z = 4.847 ***p* <.001**			0.268 (0.054) [0.162;0.374] z = 4.956 ***p* <.001**	0.268 (0.055) [0.161;0.376] z = 4.892 ***p* <.001**		
Application year					-0.010 (0.004) [-0.018;-0.002] z = -2.560 ***p* =.011**	-0.015 (0.005) [-0.024;-0.005] z = -3.009 ***p* =.003**			-0.010 (0.004) [-0.017;-0.002] z = -2.488 ***p* =.013**	-0.013 (0.005) [-0.023;-0.003] z = -2.611 ***p* =.009**
Inequality index		-0.128 (0.295) [-0.705;0.450] z = -0.433 p =.665		-0.238 (0.285) [-0.796;0.320] z = -0.835 p =.404		-0.313 (0.282) [-0.867;0.240] z = -1.110 p =.267		-0.244 (0.294) [-0.820;0.333] z = -0.830 p =.407		-0.313 (0.280) [-0.862;0.237] z = -1.115 p =.265
Study design complexity[a]		-0.014 (0.107) [-0.223;0.195] z = -0.128 p =.898		-0.007 (0.104) [-0.210;0.196] z = -0.067 p =.946		0.193 (0.126) [-0.054;0.439] z = 1.530 p =.126		-0.019 (0.109) [-0.232;0.195] z = -0.172 p =.863		0.128 (0.127) [-0.122;0.377] z = 1.002 p =.316
Proportion female authors		-0.103 (0.086) [-0.272;0.067] z = -1.188 p =.235		-0.063 (0.083) [-0.226;0.100] z = -0.760 p =.447		0.028 (0.086) [-0.140;0.195] z = 0.323 p =.747		-0.055 (0.086) [-0.222;0.113] z = -0.639 p =.523		0.029 (0.088) [-0.144;0.202] z = 0.332 p =.740
$\widehat{\sigma}_1^2$[95% CI]	0.02 [0.00;0.05]	0.02 [0.00;0.06]	0.02 [0.00;0.05]	0.02 [0.00;0.06]	0.01 [0.00;0.05]	0.01 [0.00;0.05]	0.02 [0.00;0.06]	0.03 [0.00;0.06]	0.01 [0.00;0.04]	0.01 [0.00;0.05]
$\widehat{\sigma}_2^2$[95% CI]	0.08 [0.06;0.11]	0.08 [0.06;0.11]	0.07 [0.05;0.10]	0.07 [0.05;0.10]	0.06 [0.04;0.10]	0.06 [0.04;0.10]	0.07 [0.04;0.09]	0.07 [0.04;0.09]	0.07 [0.05;0.11]	0.07 [0.05;0.11]
I^2-statistic	0.828	0.828	0.808	0.809	0.812	0.807	0.812	0.813	0.814	0.812
Q-statistic, *p*-value	1229.24 p <.0001	1202.96 p <.0001	1141.19 p <.0001	1123.13 p <.0001	819.04 p <.0001	804.91 p <.0001	1145.22 p <.0001	1127.40 p <.0001	874.90 p <.0001	864.62 p <.0001

[a] No moderators is the reference category. *Note*: The results above are based on the log odds ratio as effect size measure. P-values in bold represent the focal test of our research question.

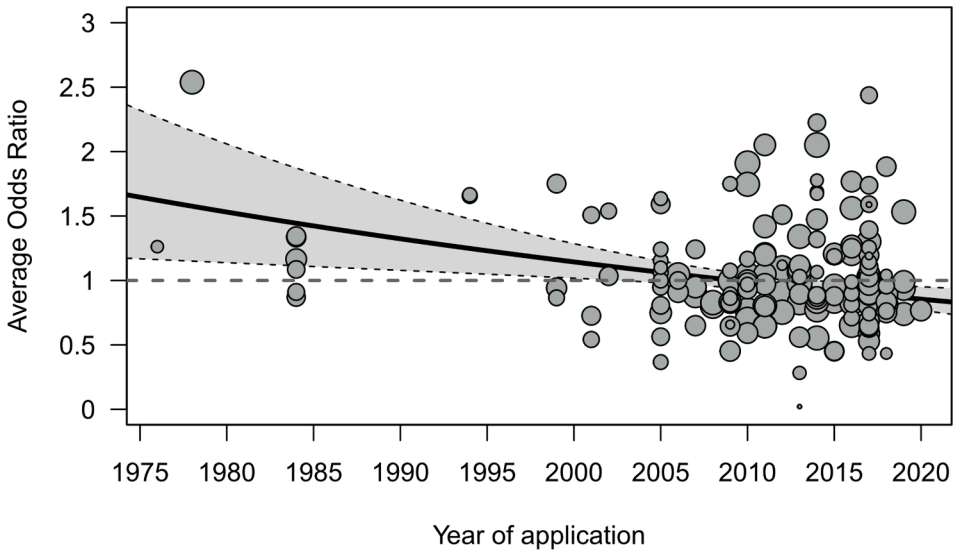


Fig. 3. Time trend of gender preferences for non-female-typed jobs. The results are based on a multilevel meta-regression model of male-typed and gender-balanced jobs combined, including gender inequality, study design complexity, and author gender ratio as control variables. Odds ratios above 1 indicate a greater preference for male applicants and odds ratios below 1 indicate greater preference for female applicants. The size of the circles is proportional to the number of applications represented by the respective data point.

3.2.4. Robustness tests

We carried out several robustness and sensitivity analyses (detailed model statistics for the analyses below are available in the R output document on the Open Science Framework at https://osf.io/ha3n4). First, three studies reported two outcomes (e.g., callbacks and interview invites) based on the same sample. This violated the assumption of independent sampling errors in the meta-analysis models (e.g., Hedges, Tipton, & Johnson, 2010). Since only six effect sizes of the 244 effect sizes came from the same sample, we decided to conduct a sensitivity analysis where we selected for each of the three studies one of the two effect sizes based on which of the two measures was more inclusive (e.g., if a study reported both callbacks and interview invites, we selected callbacks). The results of these multilevel meta-analyses did not differ substantively from those of the multilevel meta-analyses based on all data. Specifically, the overall average odds of male applicants to receive a callback remained significantly lower than the average odds of equally qualified female applicants (-0.911, $z = -3.08$, $p = .002$). The moderation of the effect of gender on callbacks by job type remained significant as well (0.253, $z = 4.82$, $p < .001$). Finally, the time trend suggesting a decrease in pro-male gender discrimination over time remained significant excluding (-0.010, $z = -2.45$, $p = .014$) or including (-0.015, $z = -3.01$, $p = .003$) control variables.

Second, although we preregistered to use publication year as the time variable, we noticed during the coding that the time between data collection and publication of an audit study varied substantially across studies (ranging from 0 to 11 years). Because the year in which applications were sent out more accurately reflects gender discrimination at any given point in time, we used data collection year as time variable in the primary analyses. However, we conducted supplemental analyses with publication year as the time variable and found comparable results. Replicating the main analyses, the time trend suggesting a decrease in pro-male gender discrimination over time remained significant excluding (-0.011, $z = -2.67$, $p = .008$) or including (-0.016, $z = -3.17$, $p = .002$) the control variables.

Third, we examined whether any one study had a particularly large effect on our results (see Section S8 in Supplementary Online Materials for detailed analyses). A leave-one-out analysis (Viechtbauer, 2010) indicated that the overall discrimination patterns and the moderation by job type remained for the most part robust. The decrease in discrimination over time was robust to the exclusion of most studies, except for one large sample study conducted in 1978 and published four years later (Firth, 1982). Exclusion of this study caused the time trend effect to be closer to zero for the models without (-0.005, $SE = 0.004$) and with control variables (-0.008, $SE = 0.006$) and made the effect nonsignificant. This is not surprising given the number of audit studies before 1990 was relatively small, such that removing the field audit with the largest sample size from the earliest time period can affect the overall estimate of the time trend (for a similar conclusion for race studies, see Quillian et al., 2017). We return to this issue in the General Discussion.

Finally, field audits aim to occlude from evaluators that they are involved in a research study by sending ostensibly real job applications to actual businesses. However, this does not completely rule out the possibility of some evaluators realizing their judgments are under scrutiny by researchers. Further, the chances of this occurring are not necessarily constant across all types of field audits. Specifically, paired audit designs may entail the greatest risk of experiment discovery among employers since they receive highly similar applications from members of both historically advantaged and underrepresented groups (e.g., men and women). However, the present meta-analysis finds no significant difference in results for audits that sent female and male applications to the same versus different employers (0.12, $SE = 0.068$, $z = 1.74$, $p = .08$).

3.3.2. Assessments of publication bias

One potential concern in meta-analyses is the presence of publication bias (Rothstein, Sutton, & Borenstein, 2006). It is possible that audit studies that document significant effects and theoretically or ideologically consistent outcomes were more likely to be published. Note that many of the currently available publication bias methods are primarily designed for univariate meta-analyses. However, Egger's regression test (Egger, Smith, Schneider, & Minder, 1997) and PET-PEESE (Stanley & Doucouliagos, 2014) are bias correcting methods that can be readily extended to multilevel meta-analysis by including the standard error of the log odds ratio as predictor in the multilevel meta-analysis with no other predictors. The other publication bias methods were applied to the univariate random-effects model where no predictors were included in the model. The included publication bias methods were: contour-enhanced funnel plots (Peters, Sutton, Jones, Abrams, & Rushton, 2008), three-parameter selection model (3PSM, Hedges & Vevea, 2005), and *p*-uniform* (van Aert, 2021). Publication bias was assessed in a

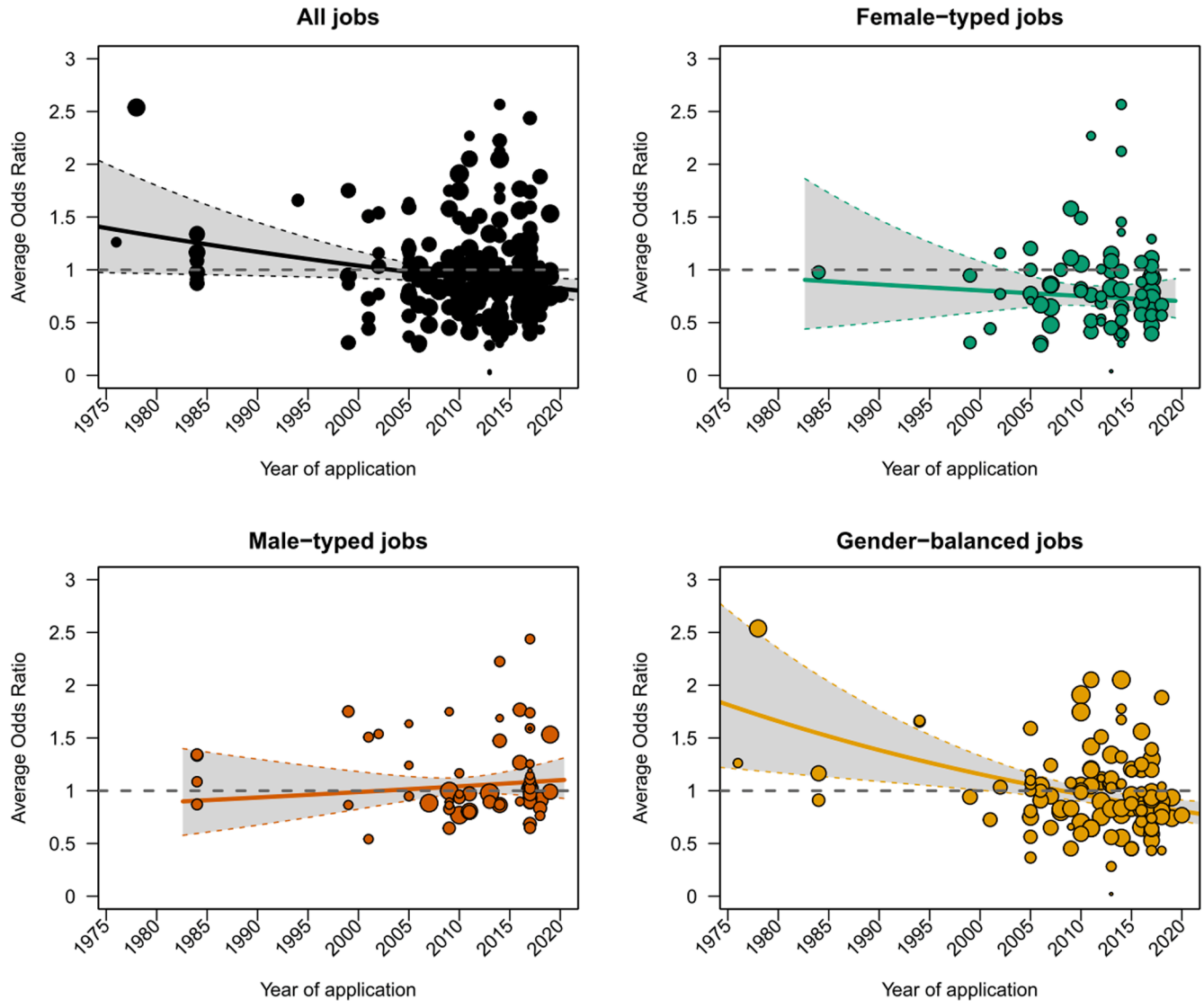


Fig. 4. Time trends of candidate gender preferences overall and by job type. The results are based on a multilevel meta-regression model including gender inequality, study design complexity, and author gender ratio as control variables. In all figures, odds ratios above 1 indicate a greater preference for male applicants and odds ratios below 1 indicate greater preference for female applicants. Error bands indicate the 95% confidence interval around the mean trend. The size of the circles is proportional to the number of applications represented by the respective data point.

meta-analysis based on all studies and based on only the studies that were published. Both assessments yielded highly comparable results, and we only report the results based on the studies that were published. The results of publication bias methods applied to all studies are available in an R output document on the Open Science Framework (https://osf.io/pt4gn).

The contour-enhanced funnel plot in Fig. 6 did not provide strong evidence for small-study effects in the meta-analysis. Further, Egger's regression test that was extended to multilevel meta-analysis was not statistically significant (-0.244, z = -0.984, p =.325). The results of the methods that correct the average log odds ratio for publication bias are presented in Table 3. The estimate of PET-PEESE that was extended to multilevel meta-analysis was slightly closer to zero (-0.032). Three different variants of the 3PSM were fitted assuming that the studies in the meta-analysis used (1) a right-tailed (2), a left-tailed, and (3) a two-tailed hypothesis test for testing the null-hypothesis of no effect. We assumed that $\alpha = 0.025$ and $\alpha = 0.05$ were used when a one-tailed and two-tailed test was conducted, respectively. The average log odds ratio was always estimated as closer to zero with 3PSM compared to the multilevel meta-analysis, and was only statistically significant in case a left-tailed hypothesis was assumed to be conducted in the studies. When applying p-uniform*, we assumed that either a right-tailed or left-tailed hypothesis test with $\alpha = 0.025$ was conducted in the studies. The average log odds ratio in both implementations of p-uniform* was closer to zero and only statistically significant in case a left-tailed hypothesis was assumed to be conducted in the studies.

Overall, the estimated average log odds ratio corrected for publication bias was closer to zero compared to the estimate of the multilevel meta-analysis. However, the combination of the non-significant Egger's regression test with a small number of statistically significant results (29.9%) suggests that there is no strong evidence for the presence of bias. This implies that if there is any publication bias in this meta-analysis, it is small—providing additional confidence in the conclusions drawn.

3.2.5. *Additional exploratory analyses*

In addition to controlling for author gender in our meta-analytic models (see above), we also examined whether author gender would moderate the extent to which a study would report gender bias on the part of prospective employers. However, author gender did not

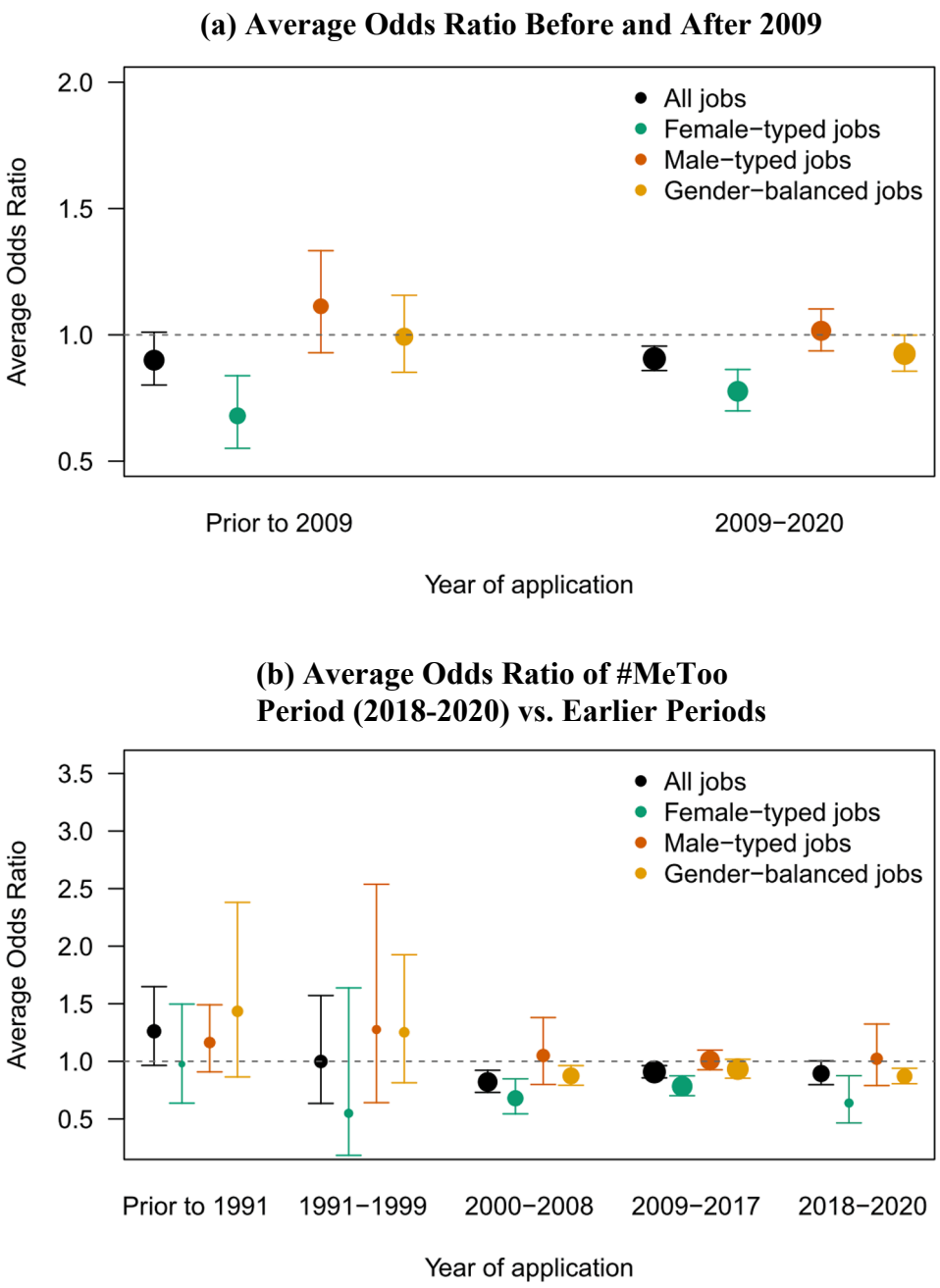


Fig. 5. Candidate gender preferences by time period and job type. Effect sizes are grouped together depending on the year the applications were sent out and were combined using a univariate random-effects model. The top panel (Fig. 5A) shows the average odds ratio before 2009 and 2009 and thereafter, which corresponds to the theoretical crossover point of the time trend in Fig. 3. The bottom panel (Fig. 5B) compares the odds ratios using more granular time periods, including the post-#MeToo years of 2018–2020. Odds ratios above 1 indicate a greater preference for male applicants and odds ratios below 1 indicate greater preference for female applicants. Error bars indicate the 95% confidence interval around the average odds ratio that is based on a normal distribution. The size of the symbols is proportional to the number of effect sizes in the respective bin.

influence the amount of gender discrimination reported across all studies and years (-0.06, $z = 0.60$, $p = .549$).

In addition to categorizing jobs into female-typed, male-typed, and gender-balanced jobs, we also examined additional job grouping variables. First, we examined whether gender discrimination would vary as a function of whether a job requires physical strength (0 = *no*, 1 = *yes*; rated by four human coders; Fleiss kappa = 0.76, $p < .001$). Results suggest that job physicality significantly moderated the effect of gender on callbacks (0.26, $z = 2.08$, $p = .038$), such that the average odds of a male (vs. female) applicant to receive a callback was significantly higher for physical jobs (odds ratio: 1.17, 95% confidence interval: 0.92, 1.49) compared to non-physical jobs (odds ratio: 0.91, 95% confidence interval: 0.85, 0.96). However, note that the proportion of jobs that require physical strength (4.92%) is small in the present sample and should thus be seen as a tentative result requiring confirmatory tests involving larger samples of jobs. Second, we explored whether gender discrimination varied as a function of whether a job required nurturance (0 = *no*, 1 = *yes*; rated by four human coders; Fleiss kappa = 0.74, $p < .001$). We found that job nurturance did not moderate the effect of gender on callbacks (0.09, $z = 0.85$, $p = .394$). Similar to job physicality, the proportion of jobs that require nurturance (6.97%) is small, rendering any conclusions tentative.

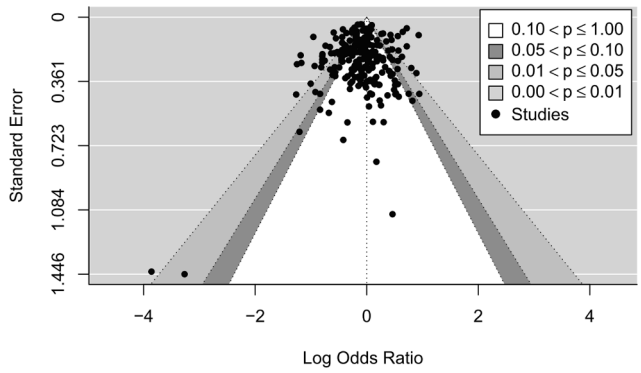


Fig. 6. Contour-enhanced funnel plot. This contour-enhanced funnel plot shows the relationship between the effect size estimates and their standard error. Shaded areas indicate the two-tailed *p*-value of a particular study. Log odds ratios above 1 indicate a greater preference for male applicants and odds ratios below 1 indicate greater preference for female applicants.

Finally, we examined the potential influence of several country-level factors that could affect gender discrimination. First, in addition to including a country's gender inequality as a control variable, we tested whether the GII (described earlier) would moderate the present results; however, this was not the case (-0.713, $z = 1.38$, $p = .168$). Second, we examined whether a country's education level would affect the results, using the United Nation's Education Index (United Nations Development Programme, 2020). Similar to the GII, we matched the closest available Education Index value to each country and study year. However, education did not moderate the effect of applicant gender on discrimination (-0.291, $z = 0.430$, $p = .667$). Third, we tested whether gender discrimination may be influenced by economic prosperity, using GDP per capita data (log) retrieved from The World Bank (https://data.worldbank.org/), but found no significant effect (-0.065, $z = 0.716$, $p = .474$). Fourth, we examined whether gender discrimination is influenced by the Human Development Index (HDI)—a summary measure of average achievement in human development, including longevity and standard of living among other factors (United Nations Development Programme, 2020). However, we found no significant moderating effect of HDI (0.036, $z = 0.077$, $p = .939$). Lastly, we examined whether culture would influence gender discrimination using the WEIRD index developed by Muthukrishna et al. (2020). However, we did not find any moderating effect (-1.465, $z = 1.417$, $p = .157$).

3.3. Discussion

In sum, we found no overall pattern of gender discrimination in hiring outcomes in favor of male applicants (Research Question 1). Based on our moderator analyses, reasons for this include that a large share of field audits in our sample were conducted in the period of 2005–2020 (see Fig. 2) and our tests of RQ1 aggregated applications for stereotypically male, gender-balanced, and female-typed jobs. Parsing the results by job type and time period indicates more favorable results for female applicants for stereotypically female jobs (Research Question 2). Further, discrimination against women for male-typed and gender-balanced jobs has diminished significantly over time (Research Question 3), although not more so in the #MeToo era than in the preceding time period (Research Question 4). At the same time, in recent years we continue to observe massive heterogeneity in discrimination-related effect size estimates across studies and settings. This suggests that there exists wide variability in current hiring practices such that discrimination against women is present in some contexts and organizations, and discrimination against men in others. At the same time, there is a reliable discriminatory bias such that male applicants for traditionally female-typed jobs (e.g., receptionist, nurse, elementary school teacher) are at a persistent disadvantage in selection decisions.

4. Study 2: Forecasting challenge

The (to us) rather surprising meta-analytic findings give rise to the related question of whether empirical patterns of gender discrimination map on to the beliefs of laypeople and academics. Accuracies and inaccuracies in perceptions of group inequalities hold important implications for the efficient allocation of limited resources to combat them (Byrd & Thompson, 2022; Ceci et al., 2023; Kraus, Hudson, & Richeson, 2022). Consider for example that gender biases in hiring may be systematically overestimated by scientists, the general public, or both. If so, workplace interventions will tend to focus on making selection contexts fairer, rather than conducting systematic audits for wage inequalities between women and men or reforming the promotion processes in organizations.

To complement the meta-analytic investigation (Study 1), we carried out an accompanying forecasting survey examining whether scientists and laypeople could accurately estimate both time-trends and the current pervasiveness of gender biases in selection settings. Previous research has demonstrated that academics sometimes perform well in anticipating the results of scientific studies, based on limited information such as article abstracts and study materials (Camerer et al., 2016; Dreber et al., 2015; Forsell et al., 2019). Accuracy on the part of scientific forecasters has been observed even for fairly complex results such as different conceptual replications testing the same research question (Landy et al., 2020), experimental designs involving complex interactions (Tierney et al., 2020; 2022), and cross-cultural similarities and differences (Tierney et al., 2021). Based on this earlier work, one straightforward prediction is that at least for academics, forecasts and realized results for gender discrimination over the years will be closely aligned.

And yet, there are also theoretical and empirical reasons to anticipate

Table 3
Overview of publication bias statistics. The table reports the parameter estimates, 95% confidence intervals, and test statistics of three publication bias metrics: PET-PEESE, three-parameter selection model (3PSM), and *p*-uniform*. Empty cells with – indicate that this result was not reported by the particular method.

		Average log odds ratio			Between-study variance (τ^2)		
		Estimate (SE)	[95% CI]	Test of no effect	Estimate (SE)	[95% CI]	Test for homogeneity
PET-PEESE		-0.032 (0.058)	[-0.146; 0.082]	$z = -0.984$, $p = .325$	–	–	–
3PSM	Right-tailed	-0.055 (0.040)	[-0.133; 0.023]	$z = -1.383$, $p = .167$	0.117 (0.021)	–	–
	Left-tailed	-0.076 (0.038)	[-0.151; -0.000]	$z = -1.722$, $p = .050$	0.097 (0.015)	–	–
	Two-tailed	-0.037 (0.050)	[-0.134; 0.061]	$z = -0.737$, $p = .461$	0.111 (0.021)	–	–
P*-uniform	Right-tailed	-0.055 (-)	[-0.124; 0.015]	$L^0 = 2.42$, $p = .120$	0.117 (-)	[0.086; 0.159]	$L^{het} = 570.000$, $p < .001$
	Left-tailed	-0.084 (-)	[–0.156; –0.013]	$L^0 = 5.36$, $p = .021$	0.098 (-)	[0.074; 0.132]	$L^{het} = 600.000$, $p < .001$

systematic inaccuracies in academics' forecasts about gender discrimination. The famous wisdom of the crowd effect (Larrick, Mannes, Soll, & Krueger, 2012; Surowiecki, 2005) relies on the removal of random noise from estimates: errors that are randomly distributed across different independent forecasters cancel each other out in the aggregate. Select and expert crowds, for example scientists relative to laypeople, should be especially accurate because their superior knowledge, skill, and strategies lead to more accurate central tendency estimates and fewer random errors (Budescu & Chen, 2015; Mannes, Soll, & Larrick, 2014). However, if different members of the crowd are systematically biased in the same direction for any reason, aggregation will fail to remove such noise from the forecasts (Lorenz, Rauhut, Schweitzer, & Helbing, 2011). The lack of political diversity among academics could represent one major source of shared systematic bias (Clark & Winegard, 2020; Duarte et al., 2015), leading scientists to overestimate the pervasiveness of gender discrimination in hiring despite their knowledge and expertise.

Another account incorporates elements of both the wisdom of the crowd and bias of the crowd predictions. It is also directly inspired by recent challenges in which scientists attempted to forecast the replicability of experimental laboratory demonstrations of gender discrimination (Tierney et al., 2022; Tierney et al., 2020). Academic forecasters were adept at anticipating not only simple condition differences, but even the results of complex designs capturing potential interactions between variables (e.g., expressions of anger or sadness by female or male targets perceived by evaluators of different genders; Tierney et al., 2022). But while the overall pattern of anticipated results mapped onto (i.e., correlated with) the replication effect sizes, in absolute terms scientists' expectations regarding overall discrimination were way off the mark. In Tierney et al. (2020), scientists predicted that Uhlmann and Cohen's (2005) findings of bias against female job candidates would emerge again in 2019, but the replication results were in the reverse direction (i.e., anti-male discrimination). Similarly, in Tierney et al. (2022), scientists expected that backlash against women who express anger in workplace settings (Brescoll & Uhlmann, 2008) would replicate, but in the new data collections the consequences of anger for perceived status, competence, likability, dominance, and assertiveness were the same for female and male targets. This leads to the prediction that academic forecasts and the realized results of field audits on gender discrimination should likewise be calibrated at a correlational level (wisdom of the crowd due to canceling out random error), but that discrimination will be overestimated in absolute terms (bias of the crowd due to systematic shared errors).

Of further interest were potential accuracies and inaccuracies among lay forecasters in this space. Even among non-academics, the widespread dissemination of classic academic studies on gender bias, some of them conducted decades ago, along with media coverage of high-profile cases of real-world discrimination, could contribute to similar systematic errors. At the same time, evidence that even laypeople can predict the results of some scientific studies (DellaVigna & Pope, 2018) and greater political diversity in the general U.S. population than among academics (Duarte et al., 2015), gives some reason to anticipate that an *inexpert* crowd of laypeople could be more collectively unbiased and accurate in this space than scientific experts. Finally, considerable evidence indicates that U.S. laypeople chronically underestimate race-based wealth inequalities (Kraus et al., 2022; Kraus, Onyeador, Daumeyer, Rucker, & Richeson, 2019). Thus, forecasts for laypeople could reflect system justifying motives (Jost & Banaji, 1994) or mere ignorance of group-based inequalities, leading to underestimations of gender gaps in hiring outcomes especially for earlier decades where the differences are larger (see Study 1).

To adjudicate between these competing possibilities, academic and lay forecasters were asked to predict the meta-analytic findings, separately for male-typed/gender-balanced and female-typed jobs, for successive spans of years. This enabled us to examine the extent to which lay and expert beliefs about the temporal trajectory and overall severity of gender discrimination map on to the observed empirical results. It further allowed us to assess correlational accuracy, absolute differences in estimated and observed effect sizes, and the potential moderating roles of forecaster characteristics. These were treated as empirical questions with multiple plausible outcomes. In other words, there were theoretical reasons to expect forecasted and realized effect sizes to correlate highly, but also weakly. Similarly, forecasts regarding absolute levels of gender discrimination might be close to the meta-analytic effect sizes or way off the mark. Moreover, either scientists or laypeople, and gender egalitarians or inegalitarians, could plausibly hold the advantage in predicting the empirical outcomes of the project.

We collected forecasts from two groups: 1) scientists primarily from the social and behavioural sciences, and 2) a nationally representative sample of laypersons from the United States. For both groups, we assessed demographic information such as their gender, political orientation on both social and economic issues, and individual differences in system-justifying vs. egalitarian beliefs (Jost & Kay, 2005; Kay & Jost, 2003). For academics, we further gathered potentially relevant disciplinary and topic expertise, such as whether they had previously published peer-reviewed research articles on gender. Greater topic expertise could enhance predictive accuracy (Budescu & Chen, 2015; Mannes et al., 2014), be associated with social-political values that increase systematic error thereby reducing accuracy (Duarte et al., 2015), or make no significant difference.

4.1. Methods

4.1.1. Forecasters

The nationally representative sample of laypeople was recruited through Prolific Academic and included 499 participants with ages between 18 and 78 (mean 35). When asked for their gender, 248 selected 'female', 244 selected 'male', 6 selected 'other', and 1 did not respond. In terms of overall political views, 85 participants reported to be at least somewhat conservative, 95 reported to be in the 'middle of the road' and 318 reported to be at least somewhat liberal. The sample was designed to be as representative as possible of the U.S. population on the dimensions of age, sex, and ethnicity using census data from the U.S. Census Bureau. Although the Prolific sample reflects the general population of the United States on these dimensions, it is not nearly as ideologically diverse as would be ideal. A sample with more left-leaning than right-leaning Americans is typical of such onsite data collection sites (Levay, Freese, & Druckman, 2016).

Forecasters from the academic sample were recruited through social media, professional listservs, direct email, and doctoral seminars. In the academic sample ($N = 312$), the age of the participants ranged from 21 to 76 (mean 38). When asked for their gender, 116 participants selected 'female', 195 selected 'male', and 1 selected 'other'. Most academics reported being at least somewhat liberal in their overall political views (247), while 38 chose 'middle of the road' and 27 reported being at least somewhat conservative. The largest subgroups of academic forecasters were from the fields of psychology (139, including subfields such as social and clinical psychology), economics (64, including subfields such as behavioural economics) and management (41, including subfields such as organizational behaviour and marketing). Of the remaining 65 participants, 35 were distributed over 16 different fields, and 33 did not provide an academic field or responded with 'N/A'. Career stages included Assistant Professor (69), Associate Professor (57), Professor (63), Graduate Student (64), Postdoctoral Scholar (27), Teaching Faculty (12), Research Assistant (11), other academic position (6), and Professor Emeritus (1); two participants did not respond to this item. Forecasters were provided a copy of the draft empirical report in advance and asked if they would like to opt-in to consortium authorship and if so to provide their names and affiliations. Colleagues listed as members of the Gender Audits Forecasting Collaboration in the Appendix A both made forecasts and indicated they would like to be part of the consortium credit. Not all forecasters elected to be listed as consortium authors, thus the number of names in the Appendix A differs

from the sample size for Study 2.

4.1.2. Materials and procedures

Instructions. Forecasters were told that a forthcoming meta-analytic investigation tested for gender biases in hiring decisions, analyzing all available studies from 1976 to 2020 in which nearly identical applications were submitted to employers by either a female candidate or a male candidate and callbacks were recorded (e.g., interview invitations, job offers). Their goal in the present survey was to try and predict the results of the meta-analytic investigation. Forecasters were provided with a link to the Study 1 methods, with results redacted.

Prediction task. The forecasters predicted the callback rates for female and male candidates, separately for female-typed jobs and for male-typed/gender-balanced jobs. Within each category, predictions were divided into four successive spans of years: 1976–1986, 1987–1997, 1998–2008, and 2009–2020. They were also asked to make an overall prediction collapsing across time periods (i.e., from 1976 to 2020). For each span of years, forecasters were presented with a column asking for "Percentage of women who received callbacks" and "Percentage of men who received callbacks". Their predictions were then converted to log odds ratios and compared to the observed log odds ratios from Study 1's meta-analysis.

System-justifying beliefs. Next, forecasters completed the general system justification scale (Kay & Jost, 2003) where high overall scores reflect a tendency to justify the existing social order ("In general, society is fair"; *1 = strongly disagree* to *7 = strongly agree*), and low scores reflect a rejection of social hierarchy and commitment to egalitarianism. They further completed the gender system justification scale (Jost & Kay, 2005), which features similar items specifically adapted to refer to gender inequality ("In general, relations between men and women are fair").

Demographics. Finally, all forecasters reported their political orientation, both overall and separately for economic and social issues (*1 = very liberal* to *7 = very conservative*), gender (female, male, other), age, and education level. Academic forecasters further indicated their academic career stage (e.g., Graduate Student, Postdoctoral Scholar, Teaching Faculty, Assistant Professor, Associate Professor, Professor), the year they received or expected to receive their PhD, their field of specialization, whether or not they currently held a tenured position, their number of publications on relevant topics (e.g., prejudice and discrimination, gender, race, and implicit bias), their total number of peer-reviewed publications, and the number of times they had taught a graduate level statistics or methods course. They further subjectively rated their proficiency in statistics relative to other academics (*1 = much lower than average* to *9 = much higher than average*), and familiarity with research on gender discrimination (*1 = not at all familiar* to *9 = extremely familiar*).

See the Open Science Framework (https://osf.io/ds6r2/) and the Supplementary Online Materials for the complete survey materials and pre-registered analysis plan for Study 2. The analyses below were pre-registered, unless explicitly otherwise noted. In contrast to Study 1's meta-analysis, for Study 2's forecasting survey we pre-registered both the traditional significance threshold of $p < .05$ and the more conservative $p < .005$ advocated by Benjamin et al. (2018). Some members of the forecasting team, and none of the meta-analysis team, are signatories to Benjamin et al. (2018), thus this was a compromise between different sub-teams of the larger project.

4.2. Results

4.2.1. Absolute levels of accuracy

Forecasted results (Study 2) are shown in Fig. 7 alongside the realized effect sizes from the meta-analysis of hiring audits (Study 1). For all forecasted log odds ratios, the mean is statistically significantly different from zero (one-sample *t*-tests) and is significantly different from the observed effects (two-sample z-tests). These *p*-values are summarized in Table S6-1 and Table S6-2 in the Supplementary Online Materials.

As seen in Fig. 7, forecasters correctly anticipated the moderating role of job stereotypicality, such that discrimination against women relative to men is comparatively greater in male-typed plus gender-balanced jobs than in female-typed jobs. A paired *t*-test was used to compare the forecasters' log odds ratios for male-typed/gender-balanced jobs for the entire time period with the log odds ratios for female-typed jobs for the entire time period. We find a statistically significant difference in both the academic sample (mean of differences: 2.16, $t(311) = 21.2$, $p < .001$, $d = 1.97$) and the layperson sample (mean of differences: 3.31, $t(498) = 29.1$, $p < .001$, $d = 2.04$).

Forecasters correctly believed that discrimination against women relative to men has decreased over time for male-typed plus gender-balanced jobs. A paired *t*-test was used to compare the forecasters' log odds ratios for male-typed/gender-balanced jobs for the first time period with the log odds ratios for last time period. We find a statistically significant decrease in both samples (mean of differences in the academic sample: 1.38, $t(311) = 19.0$, $p < .001$, $d = 1.04$; mean of differences in the layperson sample: 2.41, $t(498) = 27.1$, $p < .001$, $d = 1.20$). In addition, they incorrectly believed that discrimination against male candidates for stereotypically female-typed jobs has diminished substantially over time (mean of differences in the academic sample: -0.95, $t(311) = -11.5$, $p < .001$, $d = 0.62$; mean of differences in the layperson sample: −1.34, $t(498) = -11.6$, $p < .001$, $d = -0.57$).

At the same time, forecasters overestimated the overall degree of stereotype-consistent gender discrimination that would be observed in Study 1's meta-analysis. Testing the forecasted log odds ratios for male-typed/gender-balanced jobs against zero in a one-sample *t*-test reveals that forecasters believed that men experience more positive job application outcomes than women for male-typed plus neutral-typed jobs (academic sample: mean forecasted log odds = 1.15, $SE = 0.06$, $p < .001$; laypeople sample: mean forecasted log odds = 1.79, $SE = 0.07$, $p < .001$). A z-test comparing the mean of the forecasted log odd ratios for male/neutral typed jobs against the discrimination effect sizes from the meta-analysis further shows that forecasters overestimate the extent of discrimination for such jobs (academic sample: *z*-value = -18.98, $p < .001$; laypeople sample: *z*-value = -24.39, $p < .001$). In addition, forecasters correctly believed that women experience more positive job application outcomes than men for female-typed jobs (academic sample: mean forecasted log odds = -1.01, $SE = 0.07$, $p < .001$; laypeople sample: mean forecasted log odds = -1.52, $SE = 0.08$, $p < .001$), yet anticipated relatively greater discrimination against male candidates for such jobs than was actually observed (academic sample: $z = 8.68$, $p < .001$; laypeople sample: $z = 13.66$, $p < .001$).

They also believed that the expected overall pattern of gender discrimination has persisted into the present. Forecasters in both samples incorrectly expected that in the most recent time period (2009–2020), male candidates would receive more positive job application outcomes than women for male-typed and gender-balanced occupations (academic sample: mean forecasted log odds = 0.72, $SE = 0.05$, $p < .001$; laypeople sample: mean forecasted log odds = 1.11, $SE = 0.06$, $p < .001$), and consequently overestimated the degree of contemporary discrimination for such jobs (academic sample: $z = -13.66$, $p < .001$; laypeople sample: $z = -17.96$, $p < .001$). Forecasters in both samples correctly believed that over 2009–2020, female candidates experienced more positive job application outcomes than male candidates with regard to stereotypically female-typed jobs (academic sample: mean forecasted log odds = -0.69, $SE = 0.06$, $p < .001$; laypeople sample: mean forecasted log odds = -1.10, $SE = 0.06$, $p < .001$), yet at the same time overestimated the extent of such biases in hiring in recent years (academic sample: $z = 5.58$, $p < .001$; laypeople sample: $z = 10.23$, $p < .001$).

4.2.2. Correlational accuracy

Distinct from perceptions of absolute levels of discrimination, we examined whether there is a positive overall association between the

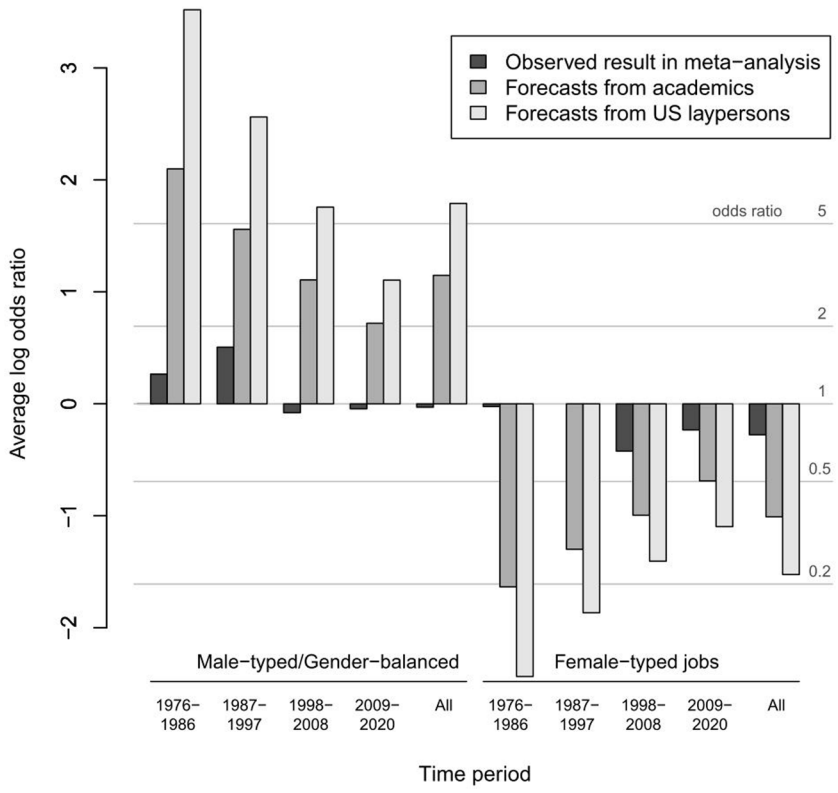


Fig. 7. Observed versus forecasted results of the gender audits meta-analysis. Observed and mean forecasted log odds ratios from the academic and U.S. nationally representative samples. Positive log odds ratios denote higher callback rates for male candidates than for female candidates, and negative log odds ratios denote higher callback rates for female candidates than for male candidates. Note that there is no observed meta-analytic result for female typed jobs for the period 1987–1997 due to a lack of relevant field audits during that span of years.

predictions of forecasters and the meta-analytic results. We test this hypothesis in an OLS regression where the individual forecast is included as an independent variable and the estimated meta-analytic gender discrimination in the forecasted time period and job type is the dependent variable. For the individual forecasts, we include the three time period predictions for female-typed jobs and the four time period predictions for male-typed/gender-balanced jobs. The forecasts for second time period for female-typed jobs is not used, because the corresponding meta-analytic effect size is missing due to a lack of audit studies during that specific span of years (see Fig. 7). We therefore have seven observations per forecaster. We include individual fixed effects in the OLS regression and we clustered standard errors at forecaster level (with the number of clusters equal to the number of forecasters) to take into account that each forecaster makes several predictions, and these predictions might be correlated. We observe a statistically significant positive correlational relationship between forecasts and observed meta-analytic outcomes for both the sample of academics (coefficient = 0.09, $t = 17.5$, $p < .001$) and the layperson sample (coefficient = 0.06, $t = 34.6$, $p < .001$). Thus, while forecasters expected much larger effects in absolute terms than emerged in the meta-analysis, there is a positive correlational relationship between their predictions and the realized results.

4.2.3. Individual differences in accuracy

Further analyses examined whether individual forecaster characteristics moderate the accuracy of their predictions. These included whether they were a trained scientist or layperson, their political orientation, and their endorsement of system justifying beliefs, among other potential moderators. Of particular interest was whether ideological beliefs, on either side of the spectrum, introduce systematic error that undermines the wisdom of the crowd effect typically observed in forecasting settings (Dreber et al., 2015).

Scientists versus laypeople. For each survey-taker, the accuracy of each forecasting question is quantified as the squared difference between the prediction and the observed estimate in the meta-analysis. We estimate the mean squared prediction error of each forecaster for the nine verifiable predictions and then test if this differs between scientists and laypeople using an independent samples *t*-test. We find that the mean squared prediction error is significantly smaller for the academic forecasters compared to the layperson sample (means = 2.86 vs. 7.27, $t(803) = -10.3$, $p < .001$). This is because laypeople gave more extreme and therefore less accurate estimates than the academics (see Fig. 7).

Political orientation. To assess the political orientation of each forecaster, we averaged their responses to the three questions about their overall, social, and economic political orientation. We estimate an OLS regression with the mean squared prediction error of each forecaster as the dependent variable and the political orientation variable as the independent variable. The OLS regression is estimated with clustered standard errors, and the test is carried out as a *t*-test on the coefficient of the political orientation variable in the OLS regression. We find no statistically significant effect of political orientation on forecasting accuracy in the academic sample (coefficient = 0.048, $t = 0.25$, $p = .80$) or in the layperson sample (coefficient = −0.33, $t = -1.52$, $p = .13$).

General system justification. For the academic sample, we find that individual differences in system justification are associated with a reduction in error (coefficient = -0.46, $t = -2.48$, $p = .014$). For the U.S. layperson sample, with increasing endorsement of the items on the general system justification scale, the error likewise decreases (coefficient = -0.83, $t = -2.65$, $p = .008$). Since a negative coefficient reflects fewer errors, this means high system justification scores are associated with greater accuracy in forecasting. Note however that these associations are only statistically significant according to the conventional $p < .05$ threshold, not under the stricter $p < .005$ threshold (Benjamin et al., 2018) we also pre-registered for the forecasting analyses (see S10

in the Supplementary Online Materials).

Gender system justification. For the academic sample, we find a statistically significant relation between individual differences in gender system justification and the accuracy of predictions (coefficient = -0.45, t = -1.98, p =.049) when the traditional p <.05 threshold is used, but not when the more conservative p <.005 threshold is employed (Benjamin et al., 2018). For the U.S. layperson sample, we observed that with increasing endorsement of the items on the gender systems justification scale, the error decreases significantly (coefficient = -1.02, t = -2.94, p =.003), regardless of which threshold is used.

Notably however, for the academic sample the association between both general and gender system justification and forecasting accuracy did not survive robustness tests (see S12 in the Supplementary Online Materials). In contrast, there is more consistent evidence that lay forecasters who were less gender egalitarian made more accurate forecasts about gender discrimination in hiring.

Topic expertise (not preregistered). We categorized forecasters who had published at least one paper on gender as a gender researcher (N = 132). A comparison group of 168 forecasters had no published work on the topic. Using an independent samples t-test, we find no statistically significant difference in the mean forecasting error between the two groups (mean of 2.70 for non-gender researchers vs. a mean of 2.90 for gender researchers, $t(250) = 0.037$, p =.71).

Forecaster gender (not preregistered). We find that accuracy differs between male and female forecasters in the U.S. representative sample, with women exhibiting a significantly higher forecasting error: 8.36 for women vs. 6.18 for men, $t(459) = 3.14$, p =.001. In the academic sample, accuracy did not statistically significantly vary with forecaster gender.

4.3. Discussion

In sum, the average forecasts from both the academic and laypeople show that they expected higher callback rates for male candidates (relative to female candidates) for male-typed/gender-balanced jobs, and higher callback rates for female candidates (relative to male candidates) for female-typed jobs. The strength of this stereotype-consistent discrimination was expected to decline from the earliest to the most recent time period, yet remain robust in recent years. Laypeople expected stronger effects compared to academics, yet both groups overestimated the severity of biases in hiring relative to the meta-analytic estimates from Study 1, most notably for the most recent time period of 2009–2020. Despite some errors in anticipating absolute levels of discrimination, a significant correlation between forecasted and realized results was observed. Scientists with a track record of publishing research on gender were not more (or less) accurate in their predictions than other academics. Some evidence emerged that less gender egalitarian laypersons were more accurate in their beliefs regarding gender biases in selection, but this effect was not robust in the academic sample and more research is needed before drawing strong conclusions. Further analyses of the forecasting results, including robustness tests, are provided in S12 in the Supplementary Online Materials.

5. General discussion

The results of Study 1′s meta-analysis of 244 effect sizes based on 85 field audits and 361,645 individual job applications across 44 years and 26 countries and territories indicate that outcomes for female candidates have become more positive over time. Until relatively recently, we observe directional preferences for men in hiring and selection for many roles. However, such discrimination against female applicants has diminished over the years in some developed societies, culminating in either no gender bias or a slight reversal in favor of female job candidates depending on the type of job and specific span of years examined. It is important to emphasize that the directional preference for female candidates that we observe in some recent time intervals are based on exploratory analyses, and was absent for stereotypically male-typed and gender-balanced jobs, where no gender bias in either direction was found.

The lack of an inflection point or sudden change in selection decisions associated with the advent of the #MeToo movement indicates that the observed decline in discrimination against women is the product of longstanding social forces rather than recent events. Returning to the question with which we opened this article, although it has been a long process, at least some societies have truly experienced meaningful change. Tellingly, however, male candidates for stereotypically female-typed jobs (e.g., secretary or elementary school teacher) did *not* receive more favorable outcomes in recent years relative to past decades. Thus, the results of the meta-analysis provide evidence of cultural stability as well as plasticity and speak to the continuing importance of gender in organizational selection decisions.

As in prior research (e.g., DellaVigna & Pope, 2018; Dreber et al., 2015), Study 2′s forecasting survey revealed a significant positive correlation between predicted and realized effect sizes for academics and a nationally representative sample of U.S. laypeople. Forecasters correctly anticipated the moderating role of the gender stereotypicality of the job (i.e., male-typed, gender-balanced, or female-typed occupations). At the same time, both groups of forecasters overestimated absolute levels of gender biases in selection decisions. Scientists predicted smaller effect sizes and were for this reason comparatively more accurate than laypeople in this regard. The forecasters correctly anticipated a decline in stereotype-consistent discrimination against female candidates since the late 1970s, but incorrectly expected that bias against male candidates for female-typed jobs would progressively diminish as well. Consistent with cultural and intellectual narratives of pervasive prejudice, both laypeople and academics believed that significant discrimination against female candidates for male-typed and gender-balanced jobs would be observed across the most recent time period (2009–2020). Scientists with higher levels of expertise in gender stereotyping, as evidenced by research publications on the topic, forecasted results for 2009–2020 along the same lines. This and other recent cases of misprediction regarding the outcomes of pre-registered tests of gender bias (Tierney et al, 2020; 2021) could result from ideological blind spots reducing forecasting accuracy in this domain. Consistent with this idea, lay forecasters who strongly rejected system justifying statements regarding gender (i.e., scored especially high in gender egalitarianism) were the least accurate at predicting the meta-analytic findings. This effect of gender system justification was conventionally statistically significant (p <.05) yet not robust to alternative analyses (see S12 in the Supplementary Online Materials) and more conservative significance cutoffs (Benjamin et al., 2018) in the academic sample. Regardless of the underlying contributors to predictive (in)accuracy, the forecasting survey indicates the meta-analytic results for recent years are profoundly counter-intuitive, even to experts, and not at all obvious based on common scientific knowledge regarding contemporary gender biases.

5.1. Mechanisms of change

Field audits are better suited to documenting the prevalence of discrimination rather than elucidating process. At the same time, that discrimination against male applicants for female-typed jobs has remained constant over the last 44 years indicates gender has not become irrelevant in contemporary workplaces. Indeed, diversity and inclusion goals, which aim to build awareness and make gender top-of-mind, may contribute to the observed cultural changes with regard to treatment of female applicants for male-typed and gender-balanced jobs. If decision makers were factoring in candidate gender less across-the-board, discrimination would have faded away across the years regardless of job type (i.e., stereotypically male, relatively gender balanced, or stereotypically female occupations and roles).

Instead, contemporary evaluators appear to be making efforts to specifically increase *female* representation in the organization, rather

than seeking to challenge stereotypes and traditional roles for both genders. Supporting this idea with independent evidence, recent research reveals muted moral concerns about male underrepresentation in traditionally female jobs, due to the perceptions that such roles are low in status and that men are not motivated to obtain them (Block et al., 2019; Reynolds et al., 2020; Stewart-Williams, Wong, Chang, & Thomas, 2022). Thus, some organizational decision makers may seek to redress a long history of discrimination and ongoing underrepresentation by supporting female candidates (Block et al., 2019; Leslie et al., 2017), yet fail to extend the same support to men whose professional interests challenge traditional gender roles.

Another likely contributor is selective shifts in stereotypes. The previously widespread belief that women are less competent than men is no longer observable in representative surveys (Eagly et al., 2020). Reductions, and in some nations full reversals, of gender gaps in education levels occurred during the period studied, eroding the motivation to engage in statistical discrimination based on perceived group differences in skills and human capital. Yet at the same time, the belief that men are less communal than women has not only failed to fade away over years, it has in fact intensified (Eagly et al., 2020). Many female-typed jobs (e.g., elementary school teacher) are perceived as communally demanding, likely contributing to ongoing discrimination against male applicants for such positions. A full elucidation of underlying mechanisms awaits more controlled laboratory investigations, for example via contemporary replications of classic gender bias experiments featuring rigorous tests of potential moderators and mediators.

5.2. Limitations and non-implications

Our most important data limitation is the comparatively smaller number of audits before 2000, and especially before 1980, compared to more recent years where more precise estimates are possible (see Quillian et al., 2017, for a similar temporal distribution of audits of racial bias). Of particular concern, Study 1′s leave-one-out analysis finds that omitting a single large early study renders the overall time trend nonsignificant. Although readers can judge for themselves, we believe a large effect size for discrimination against female job applicants in a rare well-powered study from 1978 is highly representative of the widespread discrimination against women during that period (Avent-Holt & Tomaskovic-Devey, 2012; Blau & Kahn, 1997; Eversley & Habell-Pallán, 2015; Snipp & Cheung, 2016; Stanley & Jarrell, 1998) and important data to include in the meta-analysis. The study in question features by far the largest sample from before 1980, to the point that arbitrarily deleting it from the meta-analysis excludes 93% of pre-1980 applications without any real justification. An argument must also be followed where it leads. Deleting the large 1978 study and rejecting the conclusion of a time trend necessitates also concluding little to no discrimination against female applicants for stereotypically male-typed and gender-balanced jobs prior to 1980. Rejecting the time trend also does not question another key finding: contrary to popular and scientific beliefs, there is no evidence of recent discrimination in callback rates against female job candidates in the nations sampled. If there is no time trend, both scientists and laypeople are even more inaccurate in their theories of bias against female candidates, not only misestimating present day discrimination, but also past discrimination and cultural trajectories over time as well.

Although we believe the data does support a downward time trend, pinpointing exactly when anti-female discrimination in selection decisions reached zero in the societies in question may not be possible due to data limitations. Our sample of pre-1980 observations is neither large in absolute terms nor in comparison to recent large-scale audit studies. In general, we face rapidly mounting uncertainty in meta-analyzing the literature the further we go back in time. The available set of field audits suggest that selection bias against women for male-typed and gender-balanced jobs faded away in 2009, but this conclusion may be unduly affected by one older study. The actual year could be earlier, or later, and likely differs across societies based on unmeasured moderators we are unable to capture or test due to inadequate sample sizes of older audit studies within each nation. Although drawing strong inferences about past discrimination is challenging, as discussed in greater depth below, the scientific community is well positioned to do rigorous new work testing the robustness and direction of current gender biases in selection decisions.

At the same time, we warn against interpreting our meta-analytic results to conclude equality of treatment of female applicants has been achieved with regard to historically male-typed and gender-balanced jobs, and that current efforts to increase the proportion of female employees in such roles are no longer needed. Our data did not examine the consequences of abandoning current policies, and doing so risks increasing gender bias in the future. If organizations decide to discontinue their diversity, equity, and inclusion (DEI) efforts with regard to gender, or individuals stop making the effort to override their own sexist biases, one potential result is a slide back to discrimination against qualified female applicants. A point estimate for gender discrimination close to zero in some contemporary societies also does not mean that all the industries and organizations within those societies are free of bias. It is not possible to make generic recommendations given the large heterogeneity observed in the effect sizes, and the decision to pursue inclusive hiring needs to be made on an organization-by-organization basis. Firms that experience an upward trend in hiring women may experience backlash against this increased diversity among members of historically privileged groups (Craig & Richeson, 2014; Danbold & Huo, 2015; Dover, Major, & Kaiser, 2016). Further, given the continuing discrimination against male applicants for female-typed occupations, it is important to work to improve the social acceptability and presence of men in jobs such as social worker, nursing, preschool educator, and receptionist. Even without gender bias in selection into jobs, implicit barriers remain in place that could reduce female representation in male-typed jobs and male representation in female-typed job. For example, if male nurses and secretaries are perceived to violate prescriptive gender norms and suffer backlash effects, then there should be relative fewer male applicants for such roles even in those organizations that would not have been averse to hiring them.

We find evidence of an improvement in entry-level job application outcomes for female candidates over time, as well as no overall bias against women in callback rates over the last decade. However, gender gaps may persist for other outcomes besides employer responses to initial first-round job applications. Organizations may balance their shortlists of candidates, perhaps due to DEI initiatives, and then proceed to make biased final selection decisions. Gender bias may also persist for high-level, lucrative roles, such as executive positions or elite jobs requiring specialized experience and background, for which audit studies with bogus applicants are not feasible at scale. Unfair gaps between women and men also occur across further dimensions such as wages (Auspurg, Hinz, & Sauer, 2017; Joshi, Son, & Roh, 2015; Bar-Haim, Chauvel, Gornick, & Hartung, 2018; Ceci et al., 2023; Obloj & Zenger, 2022), advancement within firms (Goldin, Kerr, Olivetti, & Barth, 2017), career penalties for parenthood (Dias, Chance, & Buchanan, 2020), and becoming the target of sexual harassment (Quick & McFadyen, 2017), among others. Even superficially gender-neutral performance criteria can create unfair gender disparities if they leave parents and caregivers at a competitive disadvantage (Cheryan & Markus, 2020). Further, the studies included in our meta-analysis examined discrimination against cisgender individuals, and transgender applicants may experience far more mistreatment on various fronts in employment settings (e.g., James et al., 2016).

Contemporary selection-stage biases against women are also probable in nations higher in gender inequality or on a different cultural trajectory (Norris & Inglehart, 2004) than those captured in the audit studies conducted to date and included in this meta-analysis. The present set of audit studies oversampled nations with relatively low levels of gender inequality by global standards (i.e., North America, Western

Europe, developed regions of Asia Pacific). The median gender inequality index of the nations included in the meta-analysis was 0.15 (25% quartile: 0.08; 75% quartile: 0.24), placing them toward the less-unequal end of the distribution with regard to leadership representation, wages, and educational attainment (i.e., the gender inequality index of the 162 tracked countries ranged from 0.03 to 0.82 between 1995 and 2019; United Nations Development Programme, 2020). Thus, the nations included in the meta-analysis were disproportionately WEIRD (Western, Educated, Industrialized, Rich, and Democratic; Henrich, Heine, & Norenzayan, 2010; Pitesa & Gelfand, 2023), because these were the places where audit studies were conducted. Although national-level inequality, development, and culture variables did not moderate the effect in our sample, we would expect to see more overall discrimination against women for stereotypically male-typed and gender-balanced jobs, fewer total female-typed jobs, and either less change or no change over time in societies with persistently strong gender roles and norms.

Even in societies where the goal to be inclusive towards women plays a major role in deliberative selection processes, concurrently operating culturally socialized stereotypes can influence judgments when the motivation or opportunity to control prejudice is weak (Banaji & Greenwald, 2013; Crandall & Eshleman, 2003; Fazio, 1990; Gaertner & Dovidio, 1986). The high observed heterogeneity in estimates across audit studies indicates that the presence and direction of gender discrimination is likely contingent on other unobserved factors. Such moderators may include evaluator motivations, candidate qualifications, job characteristics, and organizational and national culture, among others. In light of the present findings regarding moderation by job stereotypicality and related characteristics, discrimination against female candidates may persist in very strongly male-typed occupations that require physical strength, such as certain roles in construction work and manufacturing.

Another noteworthy limitation of audit studies stems from the random assignment of candidates to different professional characteristics (e.g., strength of qualifications, type of training, employment status) and demographics (e.g., gender, race, age, physical attractiveness, social class, parental status). Although this allows for tests of causality using richly detailed materials, it can render the sample of applicants non-representative of a particular labour market's actual pool of candidates. In addition, because many audit studies manipulate multiple candidate characteristics at once without clear neutral (no-information) conditions, testing simple effects of target gender in the absence of other manipulated characteristics is often not feasible. Following on previous meta-analyses of audit studies (Flage, 2018; Lippens, Vermeiren, & Baert, 2023; Koch, D'Mello, & Sackett, 2015; Quillian et al., 2017; Quillian & Lee, 2023; Zschirnt, & Ruedin, 2016), we therefore calculated the main effects of target gender across all other candidate characteristics. This allowed us to preserve the full sample of studies and carry out informative tests of job type and trends over time.

5.3. Why shifts in gender discrimination over time but not race discrimination?

One puzzling question is why a change in bias appears to have occurred for gender and selection for jobs, but not for race (e.g., Quillian et al., 2019; Quillian & Lee, 2023; Quillian et al., 2017; Rich, 2014). In some organizations a hierarchy of diversities may have emerged, such that gender is emphasized more strongly than other dimensions of inequality such as race and ethnicity, LGBTQ+ status, and socioeconomic background. Unlike sexual orientation and SES diversity, gender is perceived as observable and thus may be seen as having more signaling value. Especially in multinational firms, gender representation may be perceived by leaders as a "50–50 problem" and more straightforward to set numeric goals for than racial diversity, given the complex dynamics and varying demographics of race across societies (Sidanius & Pratto, 1999). However, gender and race are not separate dimensions of discrimination, and workers with multiple marginalized identities (e.g., Black women) can experience unique forms of mistreatment that intersects these two identities (Purdie-Vaughns & Eibach, 2008).

Recent research directly demonstrates that perceived diversity value can mediate favorable judgments of female relative to male employees (Leslie et al., 2017), and provides evidence of organizations seeking to cynically accumulate just enough members of underrepresented groups in visible positions to manage public perceptions (Chang et al., 2019; Knippen, Shen, & Zhu, 2019; Naumovska et al., 2020). However, positive evaluations of women can also result from inferences about the candidates themselves. Some evaluators engage in "belief flipping," assuming that a female candidate, having overcome barriers that her male counterparts did not face, is superior on unmeasured variables such as work motivation (Fryer, 2007). The question then arises of why such favorable inferences are not made about members of other non-prototypical groups, such as racial minorities, or if made are insufficient to overcome discriminatory biases in hiring against them (Quillian & Lee, 2023; Quillian et al., 2017; Rich, 2014).

5.4. The need for pre-registered primary investigations and replications

Unlike many academic literatures, the present set of audit studies is not characterized by an overabundance of barely significant results, or implausibly large effect size estimates from small samples. Hence, publication bias likely did not have a major impact on the results of the meta-analysis, which was confirmed when applying multiple publication bias tests. Nevertheless, publication bias methods have their own limitations (Carter, Schönbrodt, Gervais, & Hilgard, 2019; Renkewitz & Keiner, 2019; van Aert, Wicherts, & van Assen, 2016), and although the meta-analytic approach was registered in advance, the audit studies included in our meta-analysis were generally not themselves preregistered. Thus, more strictly confirmatory experiments on group-based discrimination are needed, and eventually a meta-analysis of exclusively pre-registered investigations. Further, although the present set of field audits covered a wide array of industries, companies, and organizational roles, the positions targeted were neither sampled representatively nor systematically. Future field audits should ideally define the sample space of jobs in advance, for example positions at Fortune 500 companies that have been advertised online. Since preregistration and a well-defined sample space far from eliminate all sources of research bias (Carter et al., 2019), future investigations would be particularly productive as adversarial collaborations between scholars endorsing opposing positions on the persistence of discrimination, who plan the methodology together (Clark, Costello, Mitchell, & Tetlock, 2022; Clark & Tetlock, 2022; Mellers et al., 2001).

Based on the results of the present meta-analysis, we speculate that many classic laboratory and field investigations documenting discrimination against women will no longer replicate (i.e., will yield aggregated effect size estimates close to zero) in cultural populations subject to positive change processes. To this end, our research group has recently launched a crowdsourced initiative (Klein et al., 2014) seeking to directly replicate influential experimental studies on situational and individual factors that trigger gender discrimination. Group-based discrimination represents a special case of replication since previously observed effects may not emerge in subsequent investigations due to progressive currents in the broader society (Eagly et al., 2020; Varnum & Grossmann, 2017), in addition to improvements in research practices and other common sources of non-replicability (Nelson et al., 2018). At the same time, the high heterogeneity of estimates in the present meta-analysis points to the moderation of gender discrimination by context, rather than the absence of bias. Thus, this crowd effort will focus on factors that may activate, counteract, and reverse gender biases. Discrimination against women may not be robust in baseline (control) conditions yet emerge when women self-promote and express ambition (Okimoto & Brescoll, 2010), promote diversity initiatives (Hekman, Johnson, Foo, & Yang, 2017; Rudman, Mescher, & Moss-Racusin, 2013),

are parents (Benard & Correll, 2010) or pregnant (Bragger, Kutcher, Morgan, & Firth, 2002), or are labeled feminists (Roy, Weibust, & Miller, 2009) or affirmative action hires (Heilman, Block, & Stathatos, 1997), among other potential triggers.

We hope that the upcoming years witness a broader movement to open the science of diversity and discrimination. To maximize the informational value of future investigations, we recommend researchers adopt open science best practices such as direct replication (Klein et al., 2014; Open Science Collaboration, 2015; Simons, 2014), preregistration (Wagenmakers et al., 2012), red teams (Lakens, 2020), registered reports (Chambers et al., 2015; Scheel, Schijen, & Lakens, 2021), large-scale crowdsourced data collections (Klein et al., 2014), competitive theory testing (Tierney et al., 2021), and forecasting tournaments (Dreber et al., 2015; Tetlock et al., 2014), especially those allowing for belief updating in light of new evidence (Eitan et al., 2018).

6. Conclusion

The extent to which societies have experienced meaningful changes in how women and men are treated, and whether contemporary job candidates continue to face gender discrimination, are questions of tremendous theoretical and practical importance. The present meta-analysis finds that discrimination against female applicants for jobs historically held by men has declined significantly and is no longer observable in the last decade. In contrast, bias against male applicants for female-typed jobs has remained robust and stable over the years. These results thus demonstrate both welcome declines in and the stubborn persistence of different forms of gender discrimination. Contrary to the beliefs of laypeople and academics revealed in our forecasting survey, after years of widespread gender bias in so many aspects of professional life, at least some societies have clearly moved closer to equal treatment when it comes to applying for many jobs.

Funding

Michael Schaerer benefitted from a Tier 1 Research Grant (MSS22B011) awarded by the Ministry of Education, Singapore. Michael Schaerer, Christilene du Plessis, and Eric Uhlmann further benefitted from a Tier 2 Research Grant (MM22B03) awarded by the Ministry of Education, Singapore. Daniel Lakens was supported by the Netherlands Organization for Scientific Research (NWO) VIDI Grant 452-17-013. Anna Dreber benefitted from the Jan Wallander and Tom Hedelius Foundation (grants P21-0091 and P23-0098), Knut and Alice Wallenberg Foundation and Marianne and Marcus Wallenberg Foundation (Wallenberg Scholar grant to A.D.). Cory Clark was supported by a grant from the Searle Freedom Trust. The research also benefitted from a grant by the Wharton-INSEAD Centre for Global Research awarded to Eric Uhlmann and Cory Clark, among others. Finally, Eric L. Uhlmann also received an R&D grant from INSEAD that supported this research.

CRediT authorship contribution statement

Michael Schaerer: Conceptualization, Methodology, Investigation, Visualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Christilene du Plessis:** Conceptualization, Methodology, Investigation, Visualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing. **My Hoang Bao Nguyen:** Conceptualization, Methodology, Investigation, Visualization, Project administration, Writing - original draft, Writing - review & editing. **Robbie C.M. van Aert:** Conceptualization, Methodology, Investigation, Visualization, Project administration, Writing – original draft, Writing – review & editing. **Leo Tiokhin:** Project administration, Writing – review & editing. **Daniël Lakens:** Project administration, Writing – review & editing. **Elena Giulia Clemente:** Conceptualization, Methodology, Investigation, Visualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Thomas Pfeiffer:** Conceptualization, Methodology, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Anna Dreber:** Conceptualization, Methodology, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Magnus Johannesson:** Conceptualization, Methodology, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Cory J. Clark:** Conceptualization, Methodology, Investigation, Visualization, Funding acquisition, Writing – original draft, Writing – review & editing. **Eric Luis Uhlmann:** Conceptualization, Methodology, Investigation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Preregistrations, data, syntax, and materials for Study 1 (https://osf.io/pt4gn) and Study 2 (https://osf.io/ds6r2/) are available on the Open Science Framework.

Appendix A. Names and affiliations for the Gender Audits Forecasting Collaboration

The following co-authors lent their time and expertise as contributors to the forecasting study and agreed to be credited as "Gender Audits Forecasting Collaboration" in the author string.

Ajay T. Abraham, Seattle University.

Magdalena Adamus, Centre of Social and Psychological Sciences, Slovak Academy of Sciences & Faculty of Economics and Administration, Masaryk University.

Cinla Akinci, University of St Andrews.

Federica Alberti, University of Portsmouth.

Abdelaziz M. Alsharawy, School of Public and International Affairs, Princeton University.

Shilaan Alzahawi, Stanford University, Graduate School of Business.

Frederik Anseel, University of New South Wales.

Felix Arndt, University of Guelph.

Binnur Balkan, SOFI, Stockholm University.

Ernest Baskin, Saint Joseph's University.

Carrie E. Bearden, University of California, Los Angeles.

Eric G. Benotsch, Virginia Commonwealth University.

Stefan Bernritter, King's College London

Sheila R. Black, University of Alabama, Psychology Department.

Wiebke Bleidorn, Department of Psychology, University of Zurich.

Andrew P. Boysen, University of North Carolina, Chapel Hill.

Justin P. Brienza, University of Queensland Business School.

Mitch Brown, University of Arkansas.

Stephanie E. V. Brown, ICF International.

Joshua W. Brown, Dept. of Psychological and Brain Sciences, Indiana University, Bloomington.

Jeffrey Buckley, Technological University of the Shannon, Ireland.

Brett Buttliere, Science Studies Laboratory, Centre for European Regional and Local Studies (EUROREG), University of Warsaw, Warsaw, Poland.

Nick Byrd, Stevens Institute of Technology.

Hynek Cígler, Psychology Research Institute, Faculty of Social Studies, Masaryk University.

Tabaré Capitán, Department of Economics, Swedish University of Agricultural Sciences.

Paolo Cherubini, University of Milano-Bicocca.

Shao Yuan Chong, Independent Scholar.

Esma Esen Ciftci, Anadolu University.

Chin Wen Cong, Department of Social Science, Faculty of Social Science and Humanities, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia.

Cheryl D. Conrad, Arizona State University.

Paul Conway, University of Southampton.

Elaine Costa, University of Utah.

Jolene A. Cox, Centre for Human Factors and Sociotechnical Systems, University of the Sunshine Coast, Australia.

Daniel J. Cox, University of Manchester.

Francisco Cruz, CICPSI, Faculty of Psychology, University of Lisbon.

Ian G.J. Dawson, University of Southampton.

Elif E. Demiral, Austin Peay State University.

Jaye L. Derrick, Department of Psychology, University of Houston.

Shemal Doshi, INSEAD.

Daniel J. Dunleavy, Center for Translational Behavioral Science, Florida State University.

Justin D. Durham, University of Oklahoma.

Christian T. Elbaek, Department of Management, Aarhus University.

David A. Ellis, School of Management, University of Bath, UK.

Eyal Ert, The Hebrew University of Jerusalem.

Maria Paz Espinoza, University of the Basque Country UPV/EHU.

Sascha C. Füllbrunn, Radboud University, Institute for Management Research.

Sean Fath, Cornell University, ILR School.

Remy Furrer, Center for Bioethics, Harvard Medical School.

Lenka Fiala, Department of Economics, University of Bergen.

Adrien Alejandro Fillon, University of Cyprus.

Mattias Forsgren, Department of Psychology, Uppsala University, Uppsala, Sweden.

Agapi Thaleia Fytraki, Erasmus University Rotterdam.

Francisco B. Galarza, Universidad del Pacífico, Lima, Peru.

Linnea Gandhi, The Wharton School, University of Pennsylvania.

S. Mason Garrison, Department of Psychology, Wake Forest University.

Diogo Geraldes, School of Economics and Geary Institute, University College Dublin, Ireland.

Omid Ghasemi, School of Psychology, University of New South Wales.

Biljana Gjoneska, Macedonian Academy of Sciences and Arts.

Jennifer Gothilander, School of Health, Care and Social Welfare, Mälardalen University.

Daniel Grühn, North Carolina State University.

Manuel Grieder, UniDistance Suisse, Faculty of Economics.

Dmitry Grigoryev, HSE University.

Sebastian Hafenbrädl, IESE Business School.

Georgios Halkias, Copenhagen Business School.

Roeland Hancock, Wu Tsai Institute, Yale University.

Donald A. Hantula, Temple University.

Helen C. Harton, University of Northern Iowa.

Christian P. Hoffmann, University of Leipzig.

Felix Holzmeister, Department of Economics, University of Innsbruck.

Filip Horák, Department of Constitutional Law, Faculty of Law, Charles University.

Ann-Katrin Hosch, University of Bremen.

Hirotaka Imada, University of Kent.

Konstantinos Ioannidis, University of Amsterdam & University of Birmingham.

Bastian Jaeger, Department of Social Psychology, Tilburg University.

Moritz Janas, New York University Abu Dhabi.

Bartosz Janik, Faculty of Law and Administration, University of Silesia in Katowice.

Raghabendra Pratap KC, Rollins College.

Pamela K. Keel, Department of Psychology, Florida State University.

Jared W. Keeley, Virginia Commonwealth University.

Lucas Keller, University of Konstanz.

Douglas T. Kenrick, Arizona State University.

Kim M. Kiely, University of Wollongong.

Mikael Knutsson, Linköping University.

Aleksandra Kovacheva, University at Albany, State University of New York.

Margaret Bull Kovera, John Jay College of Criminal Justice and the Graduate Center, City University of New York.

Vladislav Krivoshchekov, Department of Psychology, University of Bern.

Elizabeth J. Krumrei-Mancuso, Pepperdine University.

Danica Kulibert, Tulane University.

David Lacko, Institute of Psychology, Czech Academy of Sciences, Brno, Czech Republic.

Edward P. Lemay, Jr., University of Maryland, College Park.

Desmond W. Leung, Baruch College & The Graduate Center, CUNY.

Flora Li, Economics Experimental Lab, Nanjing Audit University, Nanjing, China.

Hause Lin, University of Regina, Massachusetts Institute of Technology.

Kyle E. Lorenzo, Fordham University.

Lorenzo Lorenzo-Luaces, Indiana University-Bloomington.

Nigel Mantou Lou, University of Victoria.

Andrey Lovakov, German Centre for Higher Education Research and Science Studies (DZHW).

Andre Luzardo, Independent Scholar.

Samuel C. MacAulay, University of Queensland.

Christopher R. Madan, University of Nottingham.

Ola Mahmoud, University of St. Gallen, School of Economics and Political Science.

Matthew C. Makel, University of Calgary.

Silvia Mari, University of Milano-Bicocca.

Diego Marino Fages, University of Nottingham.

Abigail A. Marsh, Georgetown University.

Randy J. McCarthy, Northern Illinois University.

Brett Mercier, University of Toronto.

Taciano L. Milfont, University of Waikato.

Sergio Mittlaender, FGV Direito SP & Max Planck Institute for Social Law and Social Policy.

Amanda K. Montoya, University of California, Los Angeles.

Anne Moyer, Department of Psychology, Stony Brook University.

Kristian Ove R. Myrseth, School for Business and Society, University of York.

Daniel Navarro-Martinez, Pompeu Fabra University.

Anthony J. Nelson, The Pennsylvania State University.

Levent Neyse, WZB Social Science Center Berlin; DIW, Berlin; IZA, Bonn.

Minghui Ni, Department of Psychology, Cornell University.

Paweł Niszczota, Poznań University of Economics and Business.

Lisa Norrgren, University of Gothenburg.

Natalie A. Obrecht, William Paterson University.

Tobias Otterbring, University of Agder, Kristiansand, Norway.

Zaviera A. Panlilio, University at Buffalo, SUNY.

Lora E. Park, University at Buffalo, The State University of New York.

Shiva Pauer, University of Amsterdam, The Netherlands, & Helmut Schmidt University, Germany.

Yuri G. Pavlov, Ural Federal University, Ekaterinburg, Russia; University of Tuebingen, Tuebingen, Germany.

Imre Pentek, Babes-Bolyai University.

Juan S. Pereyra, Department of Economics. FCS - UdelaR.

Patryk Perkowski, Yeshiva University, Sy Syms School of Business.

Ethan Pew, The University of Texas at Austin, McCombs School of Business.

Zehra F. Peynircioğlu, American University.

Mark V. Pezzo, University of South Florida.

Angelo Pirrone, Centre for Philosophy of Natural and Social Science, London School of Economics.

Ori Plonsky, Technion - Israel Institute of Technology.
Jonas C.C. Porfírio, Federal University of Pernambuco.
Madeleine Pownall, School of Psychology, University of Leeds, UK.
Maciej M. Próchnicki, Jagiellonian University.
John Protzko, Central Connecticut State University.
Jan P. Röer, Witten/Herdecke University.
Dobromir Rahnev, School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA.
Harry T. Reis, University of Rochester.
Kimberly Rios, Ohio University.
David L. Rodrigues, Iscte-Instituto Universitário de Lisboa, CIS-Iscte.
Priscilla Rodriguez, INSEAD.
Yefim Roth, University of Haifa.
Bradley J. Ruffle, McMaster University.
Margaret Samahita, University College Dublin.
Aishameriane Schmidt, Erasmus Universiteit Rotterdam, Tinbergen Institute and De Nederlandsche Bank.
Martin Schoemann, Technische Universität Dresden.
Philipp Schoenegger, University of St Andrews.
David C. Schwebel, University of Alabama at Birmingham.
Adrian M. Segovia, Centro Escolar University.
Jeffrey W. Sherman, University of California, Davis.
Simon Siegenthaler, University of Texas at Dallas.
Birte Siem, Leuphana University of Lüneburg.
Miroslav Sirota, University of Essex.
Eliot R. Smith, Indiana University, Bloomington.
Antonios Stamatogiannakis, IE Business School - IE University.
Steve Stewart-Williams, University of Nottingham Malaysia.
Daniel Storage, University of Denver.
Yuxin Su, SKEMA Business School.
Eli J Talbert, University of Virginia.
Andrew R. Todd, University of California, Davis.
Mirco Tonin, Free University of Bozen-Bolzano, Italy; Research Institute for the Evaluation of Public Policies (FBK-IRVAPP), Trento, Italy.
Stefan T. Trautmann, Heidelberg University.
Giovanni A. Travaglino, Institute for the Study of Power, Crime, and Society, Department of Law and Criminology, Royal Holloway, University of London.
Jo-Ann Tsang, Baylor University.
Roel van Veldhuizen, Lund University.
Michael E. W. Varnum, Arizona State University.
Alicia A. Walf, Rensselaer Polytechnic Institute.
Lukas Wallrich, Business School, Birkbeck, University of London.
Ke Wang, Harvard University.
Deborah E. Ward, Saint Joseph's University.
Christian E. Waugh, Wake Forest University.
Tobias Wingen, Institute of General Practice and Family Medicine, University Hospital Bonn, University of Bonn, Bonn, Germany.
Jan K. Woike, University of Plymouth, UK.
Conny E. Wollbrant, University of St Andrews.
Shuping Wu, INSEAD.
Keith Wylie, Emporia State University.
Qinyu Xiao, Department of Psychology, University of Hong Kong, Hong Kong SAR, China.
Sherrie Y. Xue, INSEAD.
Ofir Yakobi, NICE Actimize
Vivian Zayas, Department of Psychology, Cornell University.
Jie Zheng, Center for Economic Research, Shandong University.
Yuyang Zhong, University of California, Berkeley.
Cristina Zogmaister, Università di Milano-Bicocca.
Camille S. Zolopa, Fordham University.

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.obhdp.2023.104280.

References

** denotes study included in meta-analysis*

*Adam, B. D. (1981). Stigma and employability: Discrimination by sex and sexual orientation in the Ontario legal profession. *Canadian Review of Sociology/Revue Canadienne de Sociologie, 18*(2), 216–221.
Adams, S. M., Gupta, A., & Leeth, J. D. (2009). Are female executives over-represented in precarious leadership positions? *British Journal of Management, 20*(1), 1–12.
*Ahmad, A. (2020). When the name matters: An experimental investigation of ethnic discrimination in the Finnish labor market. *Sociological Inquiry, 90*(3), 468–496.
*Ahmed, A. M., & Lång, E. (2019). Victimized twice: A field experiment on the employability of victims. *Victims & Offenders, 14*(7), 859–874.
*Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2013). Are gay men and lesbians discriminated against in the hiring process? *Southern Economic Journal, 79*(3), 565–585.
*Albert, R., Escot, L., & Fernández-Cornejo, J. A. (2011). A field experiment to study sex and age discrimination in the Madrid labour market. *The International Journal of Human Resource Management, 22*(02), 351–375.
Alexander, A. C., & Welzel, C. (2011). Empowering Women: The Role of Emancipative Beliefs. *European Sociological Review, 27*(3), 364–384.
*Andriessen, I., Nievers, E., Dagevos, J., & Faulk, L. (2012). Ethnic discrimination in the Dutch labor market: Its relationship with job characteristics and multiple group membership. *Work and Occupations, 39*(3), 237–269.
*Arceo-Gomez, E. O., & Campos-Vazquez, R. M. (2014). Race and marriage in the labor market: A discrimination correspondence study in a developing country. *American Economic Review, 104*(5), 376–380.
Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or" would Jesse Jackson'fail'the Implicit Association Test?". *Psychological Inquiry, 15*(4), 257–278.
*Asali, M., Pignatti, N., & Skhirtladze, S. (2018). Employment discrimination in a former Soviet Union Republic: Evidence from a field experiment. *Journal of Comparative Economics, 46*(4), 1294–1309.
Auspurg, K., Hinz, T., & Sauer, C. (2017). Why should women get less? Evidence on the gender pay gap from multifactorial survey experiments. *American Sociological Review, 82*(1), 179–210.
Avent-Holt, D., & Tomaskovic-Devey, D. (2012). Relational inequality: Gender earnings inequality in U.S. and Japanese manufacturing plants in the early 1980s. *Social Forces, 91*(1), 157–180. https://doi.org/10.1093/sf/sos068
Badura, K. L., Grijalva, E., Newman, D. A., Yan, T. T., & Jeon, G. (2018). Gender and leadership emergence: A meta-analysis and explanatory model. *Personnel Psychology, 71*(3), 335–367.
Baert, S. (2018). Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005. In S. M. Gaddis (Ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance* (pp. 63–77). Cham: Springer International Publishing.
*Baert, S., & Vujić, S. (2018). Does it pay to care? Volunteering and employment opportunities. *Journal of Population Economics, 31*(3), 819–836.
*Baert, S., De Pauw, A. S., & Deschacht, N. (2016). Do employer preferences contribute to sticky floors? *ILR Review, 69*(3), 714–736.
*Baert, S., De Visschere, S., Schoors, K., Vandenberghe, D., & Omey, E. (2016). First depressed, then discriminated against? *Social Science & Medicine, 170*, 247–254.
*Baert, S., Norga, J., Thuy, Y., & Van Hecke, M. (2016). Getting grey hairs in the labour market. An alternative experiment on age discrimination. *Journal of Economic Psychology, 57*, 86–101.
*Bailey, J., Wallace, M., & Wright, B. (2013). Are gay men and lesbians discriminated against when applying for jobs? A four-city, internet-based field experiment. *Journal of Homosexuality, 60*(6), 873–894.
Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. New York, NY: Random House (Delacorte Press).
Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2004). No place for nostalgia in science: A response to Arkes and Tetlock. *Psychological Inquiry, 15*(4), 279–289.
Bar-Haim, E., Chauvel, L., Gornick, J. C., & Hartung, A. (2018). The persistence of the gender earnings gap: Cohort trends and the role of education in twelve countries. LIS Working Paper Series.
*Beauregard, J. P., Arteau, G., & Drolet-Brassard, R. (2019). Testing à l'embauche des Québécoises et Québécois d'origine maghrébine à Québec. *Recherches Sociographiques, 60*(1), 35–61.
*Becker, S. O., Fernandes, A., & Weichselbaumer, D. (2019). Discrimination in hiring based on potential and realized fertility: Evidence from a large-scale field experiment. *Labour Economics, 59*, 139–152.
Benard, S., & Correll, S. J. (2010). Normative discrimination and the motherhood penalty. *Gender & Society, 24*(5), 616–646.
Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E., Berk, R., et al. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6–10.
*Berson, C. (2012). *Does competition induce hiring equity?* Unpublished manuscript.

*Berson, C., Laouenan, M., & Valat, E. (2020). Outsourcing recruitment as a solution to prevent discrimination: A correspondence study. *Labour Economics, 64*, Article 101838.
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.
Blau, F. D., & Kahn, L. M. (1997). Swimming upstream: Trends in the gender wage differential in the 1980s. *Journal of Labor Economics, 15*(1 Part 1), 1–42.
Block, K., Croft, A., De Souza, L., & Schmader, T. (2019). Do people care if men don't care about caring? The asymmetry in support for changing gender roles. *Journal of Experimental Social Psychology, 83*, 112–131.
*Blommaert, L., Coenders, M., & Van Tubergen, F. (2014). Discrimination of Arabic-named applicants in the Netherlands: An internet-based field experiment examining different phases in online recruitment procedures. *Social Forces, 92*(3), 957–982.
Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. *Journal of Personality and Social Psychology, 55*(5), 726–737.
Boghrati, R., & Berger, J. (2023). Quantifying cultural change: Gender bias in music. *Journal of Experimental Psychology: General, 152*(9), 2591–2602.
*Booth, A., & Leigh, A. (2010). Do employers discriminate by gender? A field experiment in female-dominated occupations. *Economics Letters, 107*(2), 236–238.
Borenstein, M., & Hedges, L. (2019). Effect sizes for meta-analysis. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (Vol. 3, pp. 207–244). New York, NY: Russell Sage Foundation.
Bowen, C. C., Swim, J. K., & Jacobs, R. R. (2000). Evaluating gender biases on actual job performance of real people: A meta-analysis. *Journal of Applied Social Psychology, 30*(10), 2194–2215.
Bragger, J. D., Kutcher, E., Morgan, J., & Firth, P. (2002). The effects of the structured interview on reducing biases against pregnant job applicants. *Sex Roles, 46*(7), 215–226.
Brescoll, V. L., & Uhlmann, E. L. (2008). Can an angry woman get ahead?: Status conferral, gender, and expression of emotion in the workplace. *Psychological Science, 19*(3), 268–275. https://doi.org/10.1111/j.1467-9280.2008.02079.x
Brescoll, V. L., Uhlmann, E. L., & Newman, G. E. (2013). The effects of system-justifying motives on endorsement of essentialist explanations for gender differences. *Journal of Personality and Social Psychology, 105*(6), 891–908.
Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science, 61*(2), 267–280.
*Busetta, G., Campolo, M. G., & Panarello, D. (2020a). The discrimination decomposition index: A new instrument to separate statistical and taste-based discrimination using first-and second-generation immigrants. *International Journal of Social Economics, 47*(12), 1577–1597.
*Busetta, G., Campolo, M. G., & Panarello, D. (2020b). Weight-based discrimination in the Italian Labor Market: An analysis of the interaction with gender and ethnicity. *The Journal of Economic Inequality, 18*, 617–637.
*Busetta, G., Fiorillo, F., & Palomba, G. (2021). The impact of attractiveness on job opportunities in Italy: A gender field experiment. *Economia Politica, 38*, 171–201.
*Bygren, M., Erlandsson, A., & Gähler, M. (2017). Do employers prefer fathers? Evidence from a field experiment testing the gender by parenthood interaction effect on callbacks to job applications. *European Sociological Review, 33*(3), 337–348.
Byrd, N. (2019). What we can (and can't) infer about implicit bias from debiasing experiments. *Synthese, 198*(2), 1427–1455.
Byrd, N., & Thompson, M. (2022). Testing for implicit bias: Values, psychometrics, and science communication. *WIREs Cognitive Science.*
Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436.
*Campos-Vazquez, R. M., & Gonzalez, E. (2020). Obesity and hiring discrimination. *Economics & Human Biology, 37*, Article 100850.
Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2020). Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics, 135*(1), 269–327.
Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2021). *Gender differences in peer recognition by economists*. National Bureau of Economic Research.
Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2023). Gender gaps at the academies. *Proceedings of the National Academy of Sciences, 120*(4), Article e2212421120. https://doi.org/10.1073/pnas.2212421120
*Carlsson, M. (2011). Does hiring discrimination cause gender segregation in the Swedish labor market? *Feminist Economics, 17*(3), 71–102.
*Carlsson, M., & Eriksson, S. (2019). Age discrimination in hiring decisions: Evidence from a field experiment in the labor market. *Labour Economics, 59*, 173–183.
*Carlsson, R., Agerström, J., Björklund, F., Carlsson, M., & Rooth, D. O. (2014). Testing for backlash in hiring: A field experiment on agency, communion, and gender. *Journal of Personnel Psychology*.
Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*(2), 115–144.
Ceci, S., Ginther, D., Kahn, S., & Williams, W. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest: A Journal of the American Psychological Society, 15*(3), 75–141.
Ceci, S. J., Kahn, S., & Williams, W. M. (2023). Exploring Gender Bias in Six Key Domains of Academic Science: An Adversarial Collaboration. *Psychological Science in the Public Interest*, 15291006231163179.
Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex, 66*, A1–A2.
Chang, E. H., Milkman, K. L., Chugh, D., & Akinola, M. (2019). Diversity thresholds: How social norms, visibility, and scrutiny relate to group composition. *Academy of Management Journal, 62*(1), 144–171.
Charlesworth, T. E., & Banaji, M. R. (2022). Patterns of implicit and explicit stereotypes III: Long-term change in gender stereotypes. *Social Psychological and Personality Science, 13*(1), 14–26.
Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science, 32*(2), 218–240.
Cheryan, S., & Markus, H. R. (2020). Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review, 127*(6), 1022–1052.
Clark, C. J., Costello, T., Mitchell, G., & Tetlock, P. E. (2022). The road less traveled: Understanding adversaries is hard but smarter than ignoring them. *Journal of Applied Research in Memory and Cognition, 11*(1), 50–53.
Clark, C. J., & Tetlock, P. E. (2022). Adversarial collaboration: The next science reform. In C. Frisby, R. Redding, W. O'Donohue, & S. Lilienfeld (Eds.), *Political Bias in Psychology: Nature, Scope, and Solutions*. New York, NY: Springer.
Clark, C. J., & Winegard, B. M. (2020). Tribalism in war and peace: The nature and evolution of ideological epistemology and its significance for modern social science. *Psychological Inquiry, 31*(1), 1–22.
Coburn, K. M., & Vevea, J. L. (2016). Package 'weightr'. *Estimating Weight-Function Models for Publication Bias*.
*Correll, S. J., & Benard, S. (2007). Getting a Job: Is There a Motherhood Penalty? *American Journal of Sociology, 112*(5), 1297–1339.
Cortes, P., & Pan, J. (2018). Occupation and gender. In L. Argys, & D. Hoffman (Eds.), *The Oxford handbook of women and the economy* (pp. 425–452). New York, NY: Oxford University Press.
*Cortina, C., Rodríguez, J., & González, M. J. (2021). Mind the job: The role of occupational characteristics in explaining gender discrimination. *Social Indicators Research, 156*(1), 91–110.
Craig, M. A., & Richeson, J. A. (2014). More diverse yet less tolerant? How the increasingly diverse racial landscape affects white Americans' racial attitudes. *Personality and Social Psychology Bulletin, 40*(6), 750–761.
Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129*(3), 414–446.
Cyrus-Lai, W., Tierney, W., du Plessis, C., Nguyen, M., Schaerer, M., Giulia Clemente, E., et al. (2022). Avoiding bias in the search for implicit bias. *Psychological Inquiry, 33*(3), 203–212.
*Dahl, M., & Krog, N. (2018). Experimental evidence of discrimination in the labour market: Intersections between ethnicity, gender, and socio-economic status. *European Sociological Review, 34*(4), 402–417.
Danbold, F., & Huo, Y. J. (2015). No longer "all-American"? Whites' defensive reactions to their numerical decline. *Social Psychological and Personality Science, 6*(2), 210–218.
*Darolia, R., Koedel, C., Martorell, P., Wilson, K., & Perez-Arce, F. (2016). Race and gender effects on employer interest in job applicants: New evidence from a resume field experiment. *Applied Economics Letters, 23*(12), 853–856.
Davison, H. K., & Burke, M. J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior, 56*(2), 225–248.
DellaVigna, S., & Pope, D. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy, 126*(6), 2410–2456.
*Deming, D. J., Yuchtman, N., Abulafi, A., Goldin, C., & Katz, L. F. (2016). The value of postsecondary credentials in the labor market: An experimental study. *American Economic Review, 106*(3), 778–806.
*Deng, W., Li, D., & Zhou, D. (2020). Beauty and job accessibility: New evidence from a field experiment. *Journal of Population Economics, 33*, 1303–1341.
*Derous, E., & Ryan, A. M. (2012). Documenting the adverse impact of resume screening: Degree of ethnic identification matters. *International Journal of Selection and Assessment, 20*(4), 464–474.
*Derous, E., Ryan, A. M., & Nguyen, H. H. D. (2012). Multiple categorization in resume screening: Examining effects on hiring discrimination against Arab applicants in field and lab settings. *Journal of Organizational Behavior, 33*(4), 544–570.
*Di Stasio, V., & Larsen, E. N. (2020). The racialized and gendered workplace: Applying an intersectional lens to a field experiment on hiring discrimination in five European labor markets. *Social Psychology Quarterly, 83*(3), 229–250.
*Dias, F. A. (2020). How skin color, class status, and gender intersect in the labor market: Evidence from a field experiment. *Research in Social Stratification and Mobility, 65*, Article 100477.
Dias, F. A., Chance, J., & Buchanan, A. (2020). The motherhood penalty and the fatherhood premium in employment during covid-19: Evidence from the United States. *Research in Social Stratification and Mobility, 69*, Article 100542.
Dover, T. L., Major, B., & Kaiser, C. R. (2016). Members of high-status groups are threatened by pro-diversity organizational messages. *Journal of Experimental Social Psychology, 62*, 58–67.
Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., et al. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, 112*(50), 15343–15347.
*Drydakis, N. (2015). Sexual orientation discrimination in the United Kingdom's labour market: A field experiment. *Human Relations, 68*(11), 1769–1796.
*Drydakis, N. (2014). Sexual orientation discrimination in the Cypriot labour market. Distastes or uncertainty? *International Journal of Manpower*.
Duarte, J. L., Crawford, J. T., Stern, S., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Ideological diversity will improve psychological science. *Behavioral and Brain Sciences, 38*, e130.

*Duguet, E., Du Parquet, L., L'horty, Y., & Petit, P. (2018). Counterproductive hiring discrimination against women: evidence from a French correspondence test. *International Journal of Manpower., 39*(1), 37–50.
*Duguet, E., Gray, D., l'Horty, Y., Du Parquet, L., & Petit, P. (2020). Labour market effects of urban riots: An experimental assessment. *Papers in Regional Science, 99*(3), 787–806.
*Edo, A., Jacquemet, N., & Yannelis, C. (2019). Language skills and homophilous hiring discrimination: Evidence from gender and racially differentiated applications. *Review of Economics of the Household, 17*, 349–376.
Eagly, A., Nater, C., Miller, D., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *American Psychologist, 75*(3), 301–315.
Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin, 111*(1), 3–22.
Eagly, A. H., & Wood, W. (2012). Social role theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 458–476). Sage Publications Ltd.
Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes, & H. M. Trautner (Eds.), *The Developmental Social Psychology of Gender* (Vol. 12, pp. 123–174). New York, NY: Psychology Press.
Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629–634.
Eitan, O., Viganola, D., Inbar, Y., Dreber, A., Johannesson, M., Pfeiffer, T., et al. (2018). Is research in social psychology politically biased? Systematic empirical tests and a forecasting survey to address the controversy. *Journal of Experimental Social Psychology, 79*, 188–199.
Eversley, S., & Habell-Pallán, M. (2015). Introduction: The 1970s. *Women's Studies Quarterly, 43*(3/4), 14–30.
Fazio, R. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75–109). San Diego, CA: Academic Press.
*Fernandes, A., Huber, M., & Plaza, C. (2019). *The effects of gender and parental occupation in the apprenticeship market*. Unpublished manuscript.
Fernandez-Mateo, I., & Fernandez, R. M. (2016). Bending the pipeline? Executive search and gender inequality in hiring for top management jobs. *Management Science, 62*(12), 3636–3655.
*Firth, M. (1982). Sex discrimination in job opportunities for women. *Sex Roles, 8*, 891–901.
Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902. https://doi.org/10.1037//0022-3514.82.6.878
Flage, A. (2018). Ethnic and gender discrimination in the rental housing market: Evidence from a meta-analysis of correspondence tests, 2006–2017. *Journal of Housing Economics, 41*, 251–273.
Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology, 75*, Article 102117.
Fryer, R. G. (2007). Belief flipping in a dynamic model of statistical discrimination. *Journal of Public Economics, 91*(5–6), 1151–1166.
*Gaddis, S. M. (2013). *A matter of degrees: Educational credentials and race and gender discrimination in the labor market*. Doctoral Dissertation, University of North Carolina at Chapel Hill.
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.
Gaertner, S., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio, & S. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61–89). San Diego, CA: Academic Press.
*Galarza, F. B., & Yamada, G. (2014). Labor market discrimination in Lima, Peru: Evidence from a field experiment. *World Development, 58*, 83–94.
*Gaulke, A., Cassidy, H., & Namingit, S. (2019). The effect of post-baccalaureate business certificates on job search: Results from a correspondence study. *Labour Economics, 61*, Article 101759.
Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist, 56*(2), 109–118.
Glick, P., Zion, C., & Nelson, C. (1988). What mediates sex discrimination in hiring decisions? *Journal of Personality and Social Psychology, 55*(2), 178–186.
Goldin, C. (2006). The rising (and then declining) significance of gender. In F. Blau, M. Brinton, & D. Grusky (Eds.), *The declining significance of gender?* (pp. 67–101). New York, NY: Russell Sage Foundation.
Goldin, C., Kerr, S. P., Olivetti, C., & Barth, E. (2017). The expanding gender earnings gap: Evidence from the LEHD-2000 Census. *American Economic Review, 107*(5), 110–114.
*Gorzig, M. M., & Rho, D. (2022). The effect of the 2016 United States presidential election on employment discrimination. *Journal of Population Economics, 35*(1), 45–88.
*Granberg, M., Andersson, P. A., & Ahmed, A. (2020). Hiring discrimination against transgender people: Evidence from a field experiment. *Labour Economics, 65*, Article 101860.
Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27.
*Gringart, E., & Helmes, E. (2001). Age discrimination in hiring practices against older adults in Western Australia: The case of accounting assistants. *Australasian Journal on Ageing, 20*(1), 23–28.
Haines, E. L., Deaux, K., & Lofaro, N. (2016). The times they are a-changing… or are they not? A comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly, 40*(3), 353–363.
Hammond, M. D., Milojev, P., Huang, Y., & Sibley, C. G. (2018). Benevolent sexism and hostile sexism across the ages. *Social Psychological and Personality Science, 9*(7), 863–874.
Hebl, M., Cheng, S. K., & Ng, L. C. (2020). Modern discrimination in organizations. *Annual Review of Organizational Psychology and Organizational Behavior, 7*, 257–282.
Hedges, L., & Vevea, J. L. (2005). Selection method approaches. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, UK: Wiley.
Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65.
Heilman, M. E., Block, C. J., & Stathatos, P. (1997). The affirmative action stigma of incompetence: Effects of performance information ambiguity. *Academy of Management Journal, 40*(3), 603–625.
Heilman, M. E., & Eagly, A. H. (2008). Gender stereotypes are alive, well, and busy producing workplace discrimination. *Industrial and Organizational Psychology, 1*(4), 393–398.
Hekman, D. R., Johnson, S. K., Foo, M.-D., & Yang, W. (2017). Does diversity-valuing behavior result in diminished performance ratings for non-white and female leaders? *Academy of Management Journal, 60*(2), 771–797.
Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83.
Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558.
*Hipp, L. (2020). Do hiring practices penalize women and benefit men for having children? Experimental evidence from Germany. *European Sociological Review, 36*(2), 250–264.
Hora, S., Badura, K. L., Lemoine, G. J., & Grijalva, E. (2021). A meta-analytic examination of the gender difference in creative performance. *Journal of Applied Psychology, 107*(11), 1926–1950.
*Horváth, G. (2020). The impact of marital status on job finding: A field experiment in the Chinese labor market. *The BE Journal of Economic Analysis & Policy, 20*(4).
Inglehart, R., & Welzel, C. (2005). Gender equality, emancipative values, and democracy. In *Modernization, Cultural Change, and Democracy: The Human Development Sequence* (pp. 272–284). Cambridge: Cambridge University Press.
*Irfan & Maurer-Fazio. (2021). Discrimination at the intersection of race, gender, sexual orientation, and gender expression: A resume audit Study. *Working Paper*.
*Jackson, M. (2009). Disadvantaged through discrimination? The role of employers in social stratification. *The British Journal of Sociology, 60*(4), 669–692.
James, S., Herman, J., Rankin, S., Keisling, M., Mottet, L., Anafi, M., et al. (2016). *The report of the 2015 US transgender survey*. Washington, DC: National Center for Transgender Equality.
Johnson, C. A., & Hawbaker, K. (2018, May 25, 2018). #MeToo: A timeline of events. *Chicago Tribune*. Retrieved from http://www.chicagotribune.com/lifestyles/ct-me-too-timeline20171208-htmlstory.html.
Jones, K. P., Peddie, C. I., Gilrane, V. L., King, E. B., & Gray, A. L. (2016). Not so subtle: A meta-analytic investigation of the correlates of subtle and overt discrimination. *Journal of Management, 42*(6), 1588–1613.
Joshi, A., Son, J., & Roh, H. (2015). When can women close the gap? A meta-analytic test of sex differences in performance and rewards. *Academy of Management Journal, 58*(5), 1516–1545.
Jost, J., & Banaji, M. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology, 33*, 1–27.
Jost, J. T. (1997). An experimental replication of the depressed-entitlement effect among women. *Psychology of Women Quarterly, 21*(3), 387–393.
Jost, J. T., & Kay, A. C. (2005). Exposure to benevolent sexism and complementary gender stereotypes: Consequences for specific and diffuse forms of system justification. *Journal of Personality and Social Psychology, 88*(3), 498–509.
Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., et al. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. In A. Brief, & B. Staw (Eds.), *Research in organizational behavior* (Vol. 29, pp. 39–69). New York, NY: Elsevier.
Kay, A. C., & Jost, J. T. (2003). Complementary justice: Effects of "Poor but Happy" and "Poor but Honest" stereotype exemplars on system justification and implicit activation of the justice motive. *Journal of Personality and Social Psychology, 85*(5), 823–837.
Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr, Bahnik, S., Bernstein, M. J., et al. (2014). Theory building through replication response to commentaries on the" Many labs" replication project. *Social Psychology, 45*(4), 307–310.
Knippen, J. M., Shen, W., & Zhu, Q. (2019). Limited progress? The effect of external pressure for board gender diversity on the increase of female directors. *Strategic Management Journal, 40*(7), 1123–1150.
Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology, 100*(1), Article 128161.
Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin, 137*(4), 616–642.
*Koellinger, P. D., Mell, J. N., Pohl, I., Roessler, C., & Treffers, T. (2015). Self-employed but looking: A labour market experiment. *Economica, 82*(325), 137–161.
Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2*(1), 61–76.

Kraus, M. W., Hudson, S.-K.-T., & Richeson, J. A. (2022). Framing, Context, and the Misperception of Black-White Wealth Inequality. *Social Psychological and Personality Science, 13*(1), 4–13.
Kraus, M. W., Onyeador, I. N., Daumeyer, N. M., Rucker, J. M., & Richeson, J. A. (2019). The misperception of racial economic inequality. *Perspectives on Psychological Science, 14*(6), 899–921.
Kunst, J. R., Bailey, A., Prendergast, C., & Gundersen, A. (2019). Sexism, rape myths and feminist identification explain gender differences in attitudes toward the #MeToo social media campaign in two countries. *Media Psychology, 22*(5), 818–843.
Lakens, D. (2020). Pandemic researchers–recruit your own best critics. *Nature, 581*(7807), 121–122.
Landy, F. J. (2008). Stereotypes, bias, and personnel decisions: Strange and stranger. *Industrial and Organizational Psychology, 1*(4), 379–392.
Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., et al. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin, 156*(5), 451–479.
Larrick, R. P., Mannes, A. E., Soll, J. B., & Krueger, J. (2012). The social psychology of the wisdom of crowds. In J. Krueger (Ed.), *Social psychology and decision making* (pp. 227–242). New York, NY: Psychology Press.
Larwood, L., Szwajkowski, E., & Rose, S. (1988). Sex and race discrimination resulting from manager-client relationships: Applying the rational bias theory of managerial discrimination. *Sex Roles, 18*(1), 9–29.
*Lennon, C. (2021). How do online degrees affect labor market prospects? Evidence from a correspondence audit study. *ILR Review, 74*(4), 920–947.
Leslie, L. M., Manchester, C. F., & Dahm, P. C. (2017). Why and when does the gender gap reverse? Diversity goals and the pay premium for high potential women. *Academy of Management Journal, 60*(2), 402–432.
Levanon, A., & Grusky, D. B. (2016). The persistence of extreme gender segregation in the twenty-first century. *American Journal of Sociology, 122*(2), 573–619.
Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *SAGE Open, 6*(1), 2158244016636433.
*Li, Y. T., & Liu, J. C. E. (2021). Auditing ethnic preference in Hong Kong's financial job market: The mediation of white privilege and Hong Kong localism. *International Sociology, 36*(1), 71–90.
*Liebkind, K., Larja, L., & Brylka, A. A. (2016). Ethnic and gender discrimination in recruitment: Experimental evidence from Finland. *Journal of Social and Political Psychology, 4*(1).
Lippens, L., Vermeiren, S., & Baert, S. (2021). *The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments*. GLO Discussion Paper.
Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, 108*(22), 9020–9025.
Luo, H., & Zhang, L. (2021). Scandal, social movement, and change: Evidence from #MeToo in Hollywood. *Management Science, 68*(2), 1–19.
Mannes, A., Soll, J., & Larrick, R. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.
*Maurer-Fazio, M., & Lei, L. (2015). "As rare as a panda": How facial attractiveness, gender, and occupation affect interview callbacks at Chinese firms. *International Journal of Manpower, 36*(1), 68–85.
Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12*(4), 269–275.
*Mihut, G. (2022). Does university prestige lead to discrimination in the labor market? Evidence from a labor market field experiment in three countries. *Studies in Higher Education, 47*(6), 1227–1242.
Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474–16479.
Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., et al. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science, 31*(6), 678–701.
Nakagawa, S., & Santos, E. S. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology, 26*(5), 1253–1274.
Naumovska, I., Wernicke, G., & Zajac, E. J. (2020). Last to come and last to go? The complex role of gender and ethnicity in the reputational penalties for directors linked to corporate fraud. *Academy of Management Journal, 63*(3), 881–902.
Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534.
*Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature, 56*(3), 799–866.
*Neumark, D., Bank, R. J., & Van Nort, K. D. (1996). Sex discrimination in restaurant hiring: An audit study. *Quarterly Journal of Economics, 111*(3), 915–942.
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.
Nisbett, R. E., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the South*. Boulder, CO: Westview Press.
*Norlander, P., Ho, G. C., Shih, M., Walters, D. J., & Pittinsky, T. L. (2020). The role of psychological stigmatization in unemployment discrimination. *Basic and Applied Social Psychology, 42*(1), 29–49.
Norris, P., & Inglehart, R. (2004). *Sacred and secular: Religion and politics worldwide*. New York, NY: Cambridge University Press.
*Nunley, J. M., Pugh, A., Romero, N., & Seals, R. A. (2015). Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment. *The BE Journal of Economic Analysis & Policy, 15*(3), 1093–1125.
Okimoto, T. G., & Brescoll, V. L. (2010). The price of power: Power seeking and backlash against female politicians. *Personality and Social Psychology Bulletin, 36*(7), 923–936.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943.
*Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy, 3*(4), 148–171.
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *British Medical Journal, 372*, n71.
Paluck, E. L. (2008). What's in a norm? Sources and processes of norm change. *Journal of Personality and Social Psychology, 96*(3), 594–600.
*Patacchini, E., Ragusa, G., & Zenou, Y. (2015). Unexplored dimensions of discrimination in Europe: Homosexuality and physical appearance. *Journal of Population Economics, 28*, 1045–1073.
*Pedulla, D. S. (2016). Penalized or protected? Gender and the consequences of nonstandard and mismatched employment histories. *American Sociological Review, 81*(2), 262–289.
Pelham, B. W., & Hetts, J. J. (2001). Underworked and overpaid: Elevated entitlement in men's self-pay. *Journal of Experimental Social Psychology, 37*(2), 93–103.
Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology, 61*(10), 991–996.
*Petit, P. (2007). The effects of age and family constraints on gender hiring discrimination: A field experiment in the French financial sector. *Labour Economics, 14*(3), 371–391.
*Phillips, D. C. (2020). Do low-wage employers discriminate against applicants with long commutes? Evidence from a correspondence experiment. *Journal of Human Resources, 55*(3), 864–901.
Pitesa, M., & Gelfand, M. J. (2023). Going beyond Western, Educated, Industrialized, Rich, and Democratic (WEIRD) samples and problems in organizational research. *Organizational Behavior and Human Decision Processes*, 1–4.
Purdie-Vaughns, V., & Eibach, R. P. (2008). Intersectional invisibility: The distinctive advantages and disadvantages of multiple subordinate-group identities. *Sex Roles, 59*(5), 377–391.
*Quadlin, N. (2018). The mark of a woman's record: Gender and academic performance in hiring. *American Sociological Review, 83*(2), 331–360.
Quick, J. C., & McFadyen, M. A. (2017). Sexual Harassment: Have We Made Any Progress? *Journal of Occupational Health Psychology, 22*(3), 286–298.
Quillian, L., Heath, A., Pager, D., Midtbøen, A. H., Fleischmann, F., & Hexel, O. (2019). Do some countries discriminate more than others? Evidence from 97 field experiments of racial discrimination in hiring. *Sociological Science, 6*, 467–496.
Quillian, L., & Lee, J. J. (2023). Trends in racial and ethnic discrimination in hiring in six Western countries. *Proceedings of the National Academy of Sciences, 120*(6), Article e2212875120. https://doi.org/10.1073/pnas.2212875120
Quillian, L., & Midtbøen, A. H. (2021). Comparative perspectives on racial discrimination in hiring: The rise of field experiments. *Annual Review of Sociology, 47*, 391–415.
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.
R Core Team. (2021). R: A language and environment for statistical computing.
Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research: A comparative evaluation of six statistical methods. *Zeitschrift für Psychologie, 227*(4), 261–279.
Reynolds, T., Howard, C., Sjåstad, H., Zhu, L., Okimoto, T. G., Baumeister, R. F., et al. (2020). Man up and take it: Gender bias in moral typecasting. *Organizational Behavior and Human Decision Processes, 161*, 120–141.
*Riach, P. A., & Rich, J. (1987). Testing for sexual discrimination in the labour market. *Australian Economic Papers, 26*(49), 165–178.
*Riach, P. A., & Rich, J. (2006). An experimental investigation of sexual discrimination in hiring in the English labor market. *The BE Journal of Economic Analysis & Policy, 6*(2), 1–22.
Riach, P. A., & Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal, 112*(483), F480–F518.
Rich, J. (2014). *What Do Field Experiments of Discrimination in Markets Tell Us? A Meta Analysis of Studies Conducted Since 2000*. Available at SSRN: https://ssrn.com/abstract=2517887.
Ridgeway, C. (1991). The social construction of status value: Gender and other nominal characteristics. *Social Forces, 70*(2), 367–386.
*Rivera, L. A., & Tilcsik, A. (2016). Class advantage, commitment penalty: The gendered effect of social class signals in an elite labor market. *American Sociological Review, 81*(6), 1097–1131.
Roth, P. L., Purvis, K. L., & Bobko, P. (2012). A meta-analysis of gender group differences for measures of job performance in field studies. *Journal of Management, 38*(2), 719–739.
Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, England: John Wiley & Sons.
Roy, R. E., Weibust, K. S., & Miller, C. T. (2009). If she'sa feminist it must not be discrimination: The power of the feminist label on observers' attributions about a sexist event. *Sex Roles, 60*(5), 422–431.
Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57*(4), 743–762.
Rudman, L. A., Mescher, K., & Moss-Racusin, C. A. (2013). Reactions to gender egalitarian men: Perceived feminization due to stigma-by-association. *Group Processes & Intergroup Relations, 16*(5), 572–599.

*Ruffle, B. J., & Shtudiner, Z. E. (2015). Are good-looking people more employable? *Management Science, 61*(8), 1760–1776.
*Saeed, A., Maqsood, S., & Rafique, A. (2019). Color matters: Field experiment to explore the impact of facial complexion in Pakistani labor market. *Journal of the Asia Pacific Economy, 24*(3), 347–363.
Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science, 4*(2), 25152459211007467.
*Sherman, A., & Barokas, G. (2019). Are happy people more employable? Evidence from field experiments. *Applied Economics Letters, 26*(17), 1384–1387.
Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge, UK: Cambridge University Press.
Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76–80.
Sng, O., Neuberg, S., Varnum, M., & Kenrick, D. (2018). The behavioral ecology of cultural psychological variation. *Psychological Review, 125*(5), 714–743.
Snipp, C. M., & Cheung, S. Y. (2016). Changes in racial and gender inequality since 1970. *The Annals of the American Academy of Political and Social Science, 663*(1), 80–98. https://doi.org/10.1177/0002716215596959
Soklaridis, S., Zahn, C., Kuper, A., Gillis, D., Taylor, V., & Whitehead, C. (2018). Men's Fear of Mentoring in the# MeToo Era-What's at Stake for Academic Medicine? *The New England Journal of Medicine, 379*(23), 2270–2274.
*Spencer, N., Urquhart, M. A., & Whitely, P. (2020). Class discrimination? Evidence from Jamaica: A racially homogeneous labor market. *Review of Radical Political Economics, 52*(1), 77–95.
Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60–78.
Stanley, T. D., & Jarrell, S. B. (1998). Gender wage discrimination bias? A meta-regression analysis. *Journal of Human Resources, 33*, 947–973.
Stewart-Williams, S., Wong, X. L., Chang, C. Y. M., & Thomas, A. G. (2022). Reactions to research on sex differences: Effect of sex favoured, researcher sex, and importance of sex-difference domain. *British Journal of Psychology, 113*, 960–986.
Surowiecki, J. (2005). *The wisdom of the crowds*. New York, NY: Random House.
Talhelm, T., Zhang, X., Oishi, S., Shimin, C., Duan, D., Lan, X., et al. (2014). Large-scale psychological differences within China explained by rice versus wheat agriculture. *Science, 344*(6184), 603–608.
Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science, 23*(4), 290–295.
Tetlock, P. E., & Mitchell, G. (2009). Implicit bias and accountability systems: What must organizations do to prevent discrimination? *Research in Organizational Behavior, 29*, 3–38.
*Thomas, K. (2018). The labor market value of taste: An experimental study of class bias in US employment. *Sociological Science, 5*, 562–595.
*Thomas, K. (2021). *Intersections of race and social-class discrimination: How class signals (re)shape the racial gap in hiring*. Working paper.
Tierney, W., Cyrus-Lai, W., Hoogeveen, S., Haaf, J., Landy, J. F., et al., & Uhlmann, E. L. (2022). *Who respects an angry woman? A pre-registered re-examination of the relationships between gender, emotion expression, and status conferral*. Unpublished manuscript.
Tierney, W., Hardy, J., III, Ebersole, C. R., Viganola, D., Clemente, E. G., Gordon, M., et al. (2021). A creative destruction approach to replication: Implicit work and sex morality across cultures. *Journal of Experimental Social Psychology, 93*, Article 104060.
Tierney, W., Hardy, J. H., Ebersole, C., Leavitt, K., Viganola, D., Clemente, E., et al. (2020). Creative destruction in science. *Organizational Behavior and Human Decision Processes, 161*, 291–309.
Tosi, H. L., & Einbender, S. W. (1985). The effects of the type and amount of information in sex discrimination research: A meta-analysis. *Academy of Management Journal, 28*(3), 712–723.
Triana, M. d. C., Gu, P., Chapa, O., Richard, O., & Colella, A. (2021). Sixty years of discrimination and diversity research in human resource management: A review with suggestions for future research directions. *Human Resource Management, 60*(1), 145-204.
Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science, 16*(6), 474–480.
United Nations Development Programme. (2020). *Human Development Report 2020: The next frontier, human development and the anthropocene*. Retrieved from New York, NY: http://hdr.undp.org/.
van Aert, R. (2021). puniform: Meta-analysis methods correcting for publication bias (Version 0.2.4).
van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*(5), 713–729.
Van Den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement, 63*(5), 765–790.
Varnum, M. E., & Grossmann, I. (2016). Pathogen prevalence is associated with cultural changes in gender equality. *Nature Human Behaviour, 1*(1), 1–4.
Varnum, M. E., & Grossmann, I. (2017). Cultural change: The how and the why. *Perspectives on Psychological Science, 12*(6), 956–972.
Vial, A. C., Brescoll, V. L., & Dovidio, J. F. (2019). Third-party prejudice accommodation increases gender discrimination. *Journal of Personality and Social Psychology, 117*(1), 73–98.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638.
Walter, S., & Cook, R. (1991). A comparison of several point estimators of the odds ratio in a single 2 x 2 contingency table. *Biometrics, 47*, 795–811.
*Weichselbaumer, D. (2004). Is it sex or personality? The impact of sex stereotypes on discrimination in applicant selection. *Eastern Economic Journal, 30*(2), 159–186.
*Weisshaar, K. (2018). From opt out to blocked out: The challenges for labor market re-entry after family-related employment lapses. *American Sociological Review, 83*(1), 34–60.
Welzel, C. (2014). Evolution, empowerment, and emancipation: How societies climb the freedom ladder. *World Development, 64*, 33–51.
Williams, M. J., & Tiedens, L. Z. (2016). The subtle suspension of backlash: A meta-analysis of penalties for women's implicit and explicit dominance behavior. *Psychological Bulletin, 142*(2), 165–197.
Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2: 1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences, 112*(17), 5360–5365.
*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.
*Yemane, R. (2020). Cumulative disadvantage? The role of race compared to ethnicity, religion, and non-white phenotype in explaining hiring discrimination in the US labour market. *Research in Social Stratification and Mobility, 69*, Article 100552.
*Yemane, R., & Fernández-Reino, M. (2021). Latinos in the United States and in Spain: The impact of ethnic group stereotypes on labour market outcomes. *Journal of Ethnic and Migration Studies, 47*(6), 1240–1260.
Zenko, M. (2015). *Red Team: How to succeed by thinking like the enemy*. Basic Books.
*Zhou, X., Zhang, J., & Song, X. (2013). *Gender discrimination in hiring: Evidence from 19,130 resumes in China*. Available at http://dx.doi.org/10.2139/ssrn.2195840.
Zschirnt, E., & Ruedin, D. (2016). Ethnic discrimination in hiring decisions: A meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies, 42*(7), 1115–1134.