

Supplementary Online Materials

for

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

Table of Contents:

- S1. Study 1: Preregistered Meta-Analysis Protocol
- S2. Study 1: Deviations From Preregistered Meta-Analysis Protocol
- S3. Study 1: Coding Rubric
- S4. Study 1: Summary of Changes Made in Response to Red Team Feedback
- S5. Study 1: Robustness Test Using Univariate Random-Effects Meta-Analyses
- S6. Study 1: Sensitivity Analyses of Alternative Gender Typicality Thresholds
- S7. Study 1: Evaluation of the Normality Assumptions in the Multilevel Meta-Analysis Model
- S8. Study 1: Influential Effect Sizes
- S9. Study 1: Red Team Member Profiles
- S10. Study 2: Pre-Registered Analysis Plan for Forecasting Survey
- S11. Study 2: Forecasting Survey Items
- S12. Study 2: Detailed Report of the Forecasting Results
- S13. Supplementary References

S1. Study 1: Preregistered Meta-Analysis Protocol

The original preregistration form is available on Open Science Framework:
https://osf.io/v6gwk/?view_only=fd55ec25f77e4ff0be69adf5de451139

1. Review title.

On the trajectory of discrimination: A meta-analysis of 40 years of field experiments on gender and hiring decisions (tentative title)

2. Anticipated or actual start date.

The project started in December 2017 when some initial literature searches were conducted to get an overview of the literature and to assess the need and feasibility for a meta-analysis. Throughout 2018-2020 an initial systematic search of the literature was conducted and data from 33 studies (about 50% of the initially identified studies) were extracted to develop a coding scheme. In December 2020, we decided to adopt the “Red Team” approach and we invited feedback from an external set of gender and methods experts (the Red Team). The Red Team approach is a group of topic and methods experts designated as “devil’s advocates” charged with “finding holes and errors in ongoing work and to challenge dominant assumptions, with the goal of improving project quality” (Lakens, 2020, p. 121). As a consequence of the Red Team feedback, we have refined and revised our research question, search approach, and planned analyses.

The purpose of this analysis plan is to preregister our revised approach for the meta-analysis, which we plan to follow going forward as we conduct an entirely new search using our improved search parameters, coding scheme, and analysis plan.

The anticipated start date using the revised approach is April 22, 2021.

Lakens, D. (2020). Pandemic researchers-recruit your own best critics. *Nature*, 581, 121-121.

3. Anticipated completion date.

September 2022 (estimate).

4. Stage of review at time of this submission.

As mentioned above, we conducted a preliminary search and coded 33 studies to develop a reliable coding scheme that we presented to the Red Team for feedback. As a consequence of the feedback from the Red Team, we are starting an entirely new search using refined search parameters – which we are preregistering here.

Review stage	Started	Completed
Preliminary searches	Yes	No
Piloting of the study selection process	Yes	No
Formal screening of search results against eligibility criteria	No	No
Data extraction	Yes	No
Risk of bias (quality) assessment	No	No
Data analysis	No	No

5. Conflicts of interest.

The authors declare that they have no known conflicts of interest.

6. Review question.

Research question: "Is there a trend over time in stereotype-consistent gender discrimination in job application outcomes?"

a) General discrimination

First, we plan to test whether there is an overall main effect of candidate gender, e.g., whether discrimination against female candidates relative to male candidates emerges overall. We will also test whether discrimination is moderated by whether the job is gender-typed (male-typed, female-typed, and neutral-typed jobs), such that discrimination against female candidates emerges selecting only male-typed and neutral-typed jobs, given prior research on gender stereotyping suggests bias against female candidates should emerge only for such jobs (and potentially reverse discrimination in favor of female candidates for female-typed jobs). We hypothesized the following:

Hypothesis 1: Men experience more positive job application outcomes than women.

Hypothesis 2: The effect of gender on job application outcomes is moderated by the job's gender typicality, such that discrimination against women relative to men is stronger in male-typed and neutral-typed jobs than female-typed jobs.

b) Discrimination over time

For the tests of trends over time, we plan to compete three theoretical perspectives. Note that the theoretical predictions below apply to male-typed and neutral-typed jobs, not female-typed jobs.

Theory 1 - Persistence of bias account: This account suggests that discrimination against female candidates remains stable over time.

Theory 2 - Fading of bias account: This account implies that discrimination against female candidates is reduced over time but never reverses (i.e., there is no anti-male discrimination).

Theory 3 - Motivated liberalism account: This account predicts that discrimination against female candidates should reduce over time, such that anti-female discrimination (i.e., preference for male over female candidates) eventually turns into anti-male discrimination (i.e., preference for female over male candidates).

To test which of these four theoretical accounts most accurately describe the data, we will test the following hypothesis:

Hypothesis 3: For male-typed and neutral-typed jobs, discrimination against women relative to men in job application outcomes decreases over time.

The three theoretical accounts are summarized in the table below:

Criteria	Theoretical account		
	Persistence of bias	Fading of bias	Motivated liberalism
A Discrimination decreases over time			
B Discrimination is present in early years			
C Discrimination is present in recent years		 or 	 (reversal)

Not included above - because they are captured by the same statistical analyses - are the “no bias” account, in which there is no discrimination, and the “reverse discrimination” account, in which employers consistently prefer female over male candidates across time.

c) Secondary hypothesis

We are additionally interested in whether there is an inflection point associated with the #MeToo movement, in other words a sudden reduction in discrimination against women relative to men in the #MeToo years (i.e., research reports from 2018 and subsequent years).

d) Moderators of interest

Gender typicality: As noted earlier, we are interested in whether gender-typicality of the job (e.g., male-typed, neutral-typed, or female-typed) moderates whether an overall effect and trend in gender discrimination is observable. The discrimination is expected to be stronger in male-typed jobs and neutral-typed jobs than for female-typed jobs.

e) Control variables of interest

Gender composition of author team: We are interested in whether changes in reported discrimination over time are potentially attributable to changes in the gender composition of author teams. To test this, we will control for the proportion of female authors of the respective study.

Gender inequality index: We are interested in whether changes in discrimination trends over time are potentially attributable to changes in research populations over time. For example, if later studies are more likely to be conducted in relatively more gender-egalitarian nations, this third variable could lead to an apparent reduction in discrimination over women when no such trend exists. Thus, we will control for the gender inequality score (United Nations Development Report) of the country in which data collection took place.

Application method: We are interested in whether changes in discrimination trends over time are potentially attributable to changes in application methods. To test this, we will control for whether the application was conducted based on a resume (e.g., postal, email, online) or in person (e.g., using live actors).

Study design complexity: We are also interested in whether changes in discrimination trends over time are potentially attributable to study design complexity. For example, if later studies are more likely to include moderating variables or interventions, the overall discrimination effect size may become smaller due to such design changes. Thus, we will control for whether or not the study manipulated experimental factors beyond gender (yes/no).

7. Searches.

1) Primary database search (January 1, 1900 to April 22, 2021)

-- Web of Science:

Databases: Web of Science Core Collection

Fields: All Fields

Document type: Article

Years: All

Search string: (gender OR sex* OR woman OR women OR female*) AND (bias OR stereotyp* OR prejudic* OR discriminat*) AND ("audit stud*" OR "audit design" OR "field experiment*" OR "field stud*" OR "correspondence stud*" OR "correspondence test*" OR "correspondence audit*" OR "randomized trial*" OR "randomized experiment*" OR "randomised trial*" OR "randomised experiment*")

-- EBSCO Host:

Databases: Business Source Ultimate, EconLit, Humanities International Complete, APA PsycArticles, APA PsycInfo, SocINDEX with Full Text

Fields: Article title, abstract, keywords

Document type: Article

Years: All

Search string: (gender OR sex* OR woman OR women OR female*) AND (bias OR stereotyp* OR prejudic* OR discriminat*) AND ("audit stud*" OR "audit design" OR "field experiment*" OR "field stud*" OR "correspondence stud*" OR "correspondence test*" OR "correspondence audit*" OR "randomized trial*" OR "randomized experiment*" OR "randomised trial*" OR "randomised experiment*")

2) Supplementary search strategy (January 1, 1900 to April 22, 2021)

-- Google Scholar

Scope: First 1,000 results

Search string: gender|sex|woman|women|female

bias|stereotyping|stereotype|prejudice|prejudiced|discrimination|discriminate

audit|field|correspondence

-- Backward citation tracking

Reviewing references of key reviews and meta-analyses

-- Forward citation tracking

Reviewing citations of top 5 cited articles from primary database search

-- Identification of unpublished studies

Post requests for unpublished studies on listservs of relevant conferences (e.g., Academy of Management, SPPS), as well as relevant sites on social media (e.g., PsychMAP and PsychMethods discussion groups on Facebook)

Emailing first authors of published studies in our sample to request unpublished studies

Search of unpublished dissertations (ProQuest Dissertations & Theses Global)

Depending on the duration of the project and peer-review process, as well as the possibility of further feedback from the Red Team, we may update our search after April 22, 2021 accordingly.

8. Types of study to be included.

We will include audit studies, correspondence studies, and field experiments that examine the effect of applicant gender on job application outcomes. The specific criteria are as follows:

- Gender is manipulated either in a “within” design (employer receives both male and female applications) or a “between” design (employer only receives either male or female applications and gender of target varies across employers). We will include both resume audit studies (in which indicators of candidate gender such as names or photos are randomly assigned to otherwise similar resumes) and in-person audit studies (in which pairs of trained testers who differ on the basis of gender but are otherwise similar apply for jobs).
- The study examines actual job application outcomes (e.g., request for additional information, callbacks, interview invites, job offers)
- The study reports sufficient information to calculate effect sizes

9. * Data extraction (selection and coding).

Study selection

- One coder will review the titles and abstracts of the final search results resulting from our primary and supplementary search strategies to determine whether the study is relevant (“Include”) or not (“Exclude”). Studies that are labelled as unclear (“Maybe”) will be reviewed by a second coder who will make the final inclusion/exclusion decision in discussion with the other coders. This review will be conducted in Rayyan (<https://rayyan.ai/>).
- Studies that are deemed as relevant based on the title and abstract review, will be reviewed in full by two coders. This review will be conducted in Rayyan. Studies that are labelled as unclear (“Maybe”) will be reviewed by a third coder who will make the final inclusion/exclusion decision in discussion with the other coders.

Data extraction

- For factual data (e.g., journal, publication year, author gender, etc.) one author will extract all data. Factual data will be cross-checked by a research assistant to identify any errors. For subjective data (e.g., gender typicality of job), two authors will independently code all studies, based on which we will calculate intercoder reliability. In case of disagreement, a third author will help make the final decision in discussion with the other two authors.
- In the case of missing data, we will contact the authors. If they are unable to provide the necessary data, we either treat the data as missing (e.g., for control variables) or exclude the study from the meta-analysis entirely when essential information is missing (e.g., for gender breakdowns, job outcomes, etc.).

Variables to be extracted or calculated

- Paper title
- Journal name
- Search source (e.g., Web of Science, Google Scholar, listserv, etc.)
- Publication Status (published vs. unpublished)
- Author names
- # female authors
- # male authors
- Proportion of female authors
- Publication year
- Year(s) during which job applications were sent out
- Single year(s) during which most job applications were sent out (e.g., middle year if 3 year period, longer period if 2 years, etc.)
- Country/countries in which data was collected
- Gender inequality index (retrieved from United Nations Human Development Report)

- Job application method: Resume audit (e.g., email, post, online form) vs. in person audit (e.g., with actors)
- Matched pair study (yes vs. no)
- Type of job that was applied for
- Is the job stereotypically male-typed, neutral-typed, or stereotypically female-typed?
- Applicant education (if reported)
- Operationalization of dependent measure (e.g., job offer, call back, interview)
- Study design complexity (other manipulations apart from gender?)
- Condition name if additional factors manipulated other than male vs. female
- Proportion of female applicants that succeeded
- Total number of female applications
- Proportion of male applicants that succeeded
- Total number of male applications
- # of female applicants that succeeded
- # of female applicants that did not succeed
- # of male applicants that succeeded
- # of male applicants that did not succeed
- Effect size based on log odds
- Standard Error based on log odds
- Total sample size (male + female)

Any deviations from the above protocol (e.g., additional variables coded after preregistration) will be highlighted in the methods section and/or the supplemental online materials of the article.

10. Risk of bias (quality) assessment.

We will study the impact of small-study effects and publication bias in a random-effects model based on all studies. We will create a contour-enhanced funnel plot (Peters, Sutton, Jones, Abrams, & Rushton, 2008) to assess whether there is evidence for small-study effects. We will test for small-study effects using Egger's regression test (Egger, Smith, Schneider, & Minder, 1997), and correct for it using PET-PEESE (Stanley & Doucouliagos, 2014). Publication bias will be studied by applying the three-parameter selection model as proposed by Vevea and Hedges (1995) and implemented in the R package "weightr" (Coburn & Vevea, 2016) and *p*-uniform* (van Aert, 2021). For Egger's regression test and the publication bias test in the three-parameter selection model, we will report the test statistic and corresponding *p*-value. For methods that correct the effect size for small-study effects and publication bias, we will report (1) the average effect size and between-study variance estimate corrected for bias, corresponding sampling error, and 95% confidence interval and (2) test-statistic and *p*-value for testing whether the average effect size corrected for bias is different from zero. Two-tailed hypothesis tests will be conducted using $\alpha=.05$.

11. Strategy for data synthesis.

We will compute for each study a log odds ratio after adding 0.5 to all cells to avoid division by zero when computing the log odds ratio and the corresponding sampling variance and to decrease bias in the estimator of the log odds ratio (Walter & Cook, 1991). Random-effects meta-regression models will be used in all our analyses since these enable to draw inferences to the population of studies. The restricted maximum likelihood estimator will be used as estimator of the between-study variance in true effect size. Two-tailed hypothesis tests with $\alpha=.05$ and 95% confidence intervals will be computed after applying the Knapp-Hartung adjustment (Hartung & Knapp, 2001; Sidik & Jonkman, 2002) to take uncertainty in the estimated within-study and between-study variances into account.

Hypothesis 1 will be tested by computing what the average effect size is after controlling for the variables listed as control variables (see section 6 "Review Question" for a list of the control variables). Hypothesis 2 will be tested by including the dichotomous moderator variable "Gender typicality" (0=

female-typed, 1= everything else) in the model. Hypothesis 3 will be tested by first creating a subset of studies with male-typed and neutral-typed jobs. A random-effects meta-regression model will then be applied using the studies in this subset with publication year and the control variables as independent variables in the model. The variable publication year will be centered at the publication year of the oldest study included in the meta-analysis to avoid convergence issues.

All analyses will be conducted in R (R Core Team, 2021) using the R package “metafor” (Viechtbauer, 2010). For each applied meta-regression model and tested hypothesis, we will report (1) estimates of meta-regression coefficients, corresponding sampling error, and 95% confidence interval, (2) test-statistic and p -value, (3) estimate of the residual between-study variance in true effect size together with a 95% confidence interval obtained with the Q -profile method (Viechtbauer, 2007), (4) the I^2 -statistic to quantify the residual between-study variance together with its confidence interval, and (5) the Q -statistic and p -value to test for residual between-study variance in the true effect sizes.

12. Language.

English

S2. Study 1: Deviations From Preregistered Meta-Analysis Protocol

Although we preregistered all steps of the meta-analysis (Study 1)—including article search, data extraction, and analyses—in some cases we had to deviate from the preregistered protocol. Following the recommendations by Nosek and colleagues (Nosek, Beck, Campbell, Flake, Hardwicke, Mellor et al., 2019; Nosek, Ebersole, DeHaven, & Mellor, 2018), we transparently report all such deviations in the table below. Specifically, the table reports the originally preregistered approach, the revised approach, as well as a justification for the respective deviation. All other aspects of the preregistration were followed unless explicitly indicated.

Table S1. Deviations from the preregistered protocol.

Phase	Preregistered Approach	Revised Approach	Justification
Search	We preregistered to solicit unpublished studies from first authors of the studies in our sample.	We solicited unpublished studies from corresponding authors of the studies in our sample.	We initially assumed that first authors would also be corresponding authors, but this was not always the case. We thought that we would obtain a higher response rate if we contacted corresponding authors.
	We preregistered to search article title, abstract, keywords on EBSCO.	We searched all fields on EBSCO.	Different databases searched via EBSCO have different functionalities in respect to search fields. Thus, we used a more generous approach than originally preregistered to avoid any issues related to such functionality differences.
	We preregistered to search for articles on the databases accessed via EBSCO.	We searched for articles on Business Source Ultimate, PsycArticles, PsycInfo, and SocINDEX. For EconLit and Humanities International Complete all document types were searched.	EconLit and Humanities International Complete do not allow restricting document type to article when searched via EBSCO. We did not anticipate this during the preregistration.
	The preregistered search string omitted an “OR” (i.e., OR “correspondence test*” “correspondence audit*” OR).	The updated search string we used included an “OR” (i.e., OR “correspondence test*” OR “correspondence audit*” OR).	This was a typo and was done correctly for the search reported in this report.
Data Extraction	Studies deemed relevant based on title and abstract, will be screened in by two coders. Ambiguities will be resolved by a 3rd coder.	Studies that were deemed relevant based on title and abstract, were screened by one coder. Ambiguities were resolved by a 2nd coder.	There were very few ambiguities in respect to inclusion/exclusion at the initial screening stage, thus screening by one author was sufficient and a second author was consulted in case of ambiguities.
	Subjective variables (i.e., gender typicality of job) will be coded by two coders and disagreements will be	Subjective variables (i.e., gender typicality of job) will be coded by four coders and disagreements will be resolved through a	During the peer review process, reviewers pointed out that coders may have cultural biases based on their origins. Thus, we doubled the number of human coders to four. The coders represent most major geographies, including South and North

	resolved through a third coder.	majority rule (in case of a 3-1 split) or through discussion (in case of a 2-2 split).	America, Europe, Africa, and Southeast Asia.
	We preregistered to code whether a study is a matched pairs study (yes vs. no)	We coded whether a study manipulated gender within- or between-employers , and whether a study matched real applicants rather than equivalent fictitious CVs.	We decided to extract two separate variables that are together more meaningful for our research question and could be coded in a more precise and granular way.
	We preregistered to exclude non-English articles .	We included non-English articles .	To address reviewer feedback, we screened 11 articles that were published in a language other than English. One such article met our inclusion criteria and was subsequently included in the meta-analysis.
Analysis	We preregistered to use the publication year of a study as the time variable to test changes in discrimination across years.	In our main analyses, we used the data collection year (i.e., the year in which most applications were sent out) to test changes in discrimination across years, and report the analysis with publication year in a robustness test.	During the coding it became apparent that there is a wide range in the amount of time it took for a study to go from data collection to publication (ranging from 0 to 11 years). We reasoned that data collection year more accurately reflects gender discrimination at any given point in time and that publication year may distort our trend analyses. However, we report a robustness test with publication year as time variable and find comparable results.
	In the prediction section, we preregistered our control variables for the time trend analysis only, while due to an oversight in the analysis section we preregistered the control variables for all analyses.	In the main manuscript, we report the analyses both with and without the control variables.	The preregistration contained mixed statements about the control variables. Our intention was to use controls to account for alternative explanations in the trend analysis. The controls are less relevant for static analyses since they represent trend-contingent factors. E.g., a disproportionate number of audits in more gender-egalitarian countries in later time periods could confound analyses of trends in discrimination across time. However, controlling for gender egalitarianism of nation is less relevant to calculating aggregate gender discrimination regardless of year and country. However, for completeness, we still report these analyses both with and without control variables.
	We preregistered to include application method as a control variable.	We did not include application method as a control variable.	There was insufficient variance for this variable. Submitting a resume was used for most effect sizes as application method (234) and only four effect sizes were based on another application method, which meant there was insufficient data to include application method as a control variable.
	We preregistered our analyses using univariate meta-regression .	We conducted our analyses using multilevel regression , but we also report the univariate models.	Because some studies contributed more than one effect size, it was more appropriate to use a multilevel approach.

	We preregistered to code gender typicality of job using two human coders	We report results of two coding approaches in the main text: 1) We use a human coding approach using four coders 2) We use a data-driven coding approach using country-level employment statistics	During the review process, the reviewers suggested that the preregistered coding approach may be subject to bias and does not take into account shifts in gender typicality over time. We therefore decided to increase the number of human coders from two to four coders representing many major world geographies (South and North American, South African, European, and East Asian). We also followed the review team's suggestion to use an exploratory, data-driven approach using country-level statistics. We expand on the advantages and disadvantages of each approach in the main text.
	We did not preregister testing for moderation by additional study, job, or country-level characteristics.	We now report exploratory analyses testing additional moderators, including author gender, job physicality, job nurturance, gender inequality, education index, GDP per capita, human development, and culture (i.e., WEIRD index)	These exploratory analyses were conducted during the revision to address reviewer questions.

S3. Study 1: Coding Rubric

The table below provides an overview of the information extracted and the coding scheme used.

Table S2. Coding Rubric.

Variable Name	Description	Coding Scheme (if any)
UniqueID	Unique identifier for each data point	-
StudyID	Study identifier	-
StudyNum	Study number in paper	-
EffectID	Identifier of effect within a particular study	-
StudyLevelData	Indicates observations to be used for study-level analyses (i.e., one observation per study)	1 = study level observation; 0 = otherwise
EffectsLevelData	Indicates observations to be used for effects-level analyses (i.e., multiple observations per study, if applicable)	1 = effects level observation; 0 = otherwise
MultipDVs	Study includes multiple dependent variables (non-independence of variance)	1 = yes; 0 = no
Title	Paper title	-
Journal	Journal name	-
Source	Search source	Web of Science, Google Scholar, etc.
Published	Is the study published or in press?	1 = yes; 0 = no
Authors	Author names	-
Email	Email address of corresponding author	-
AuthFemale	Number of female authors	-
AuthMale	Number of male authors	-
AuthFemProp	Proportion of female authors	-
PubYear	Publication year	-
ApplicYear	Year(s) of job applications	-
ApplicYearMost	Single year(s) during which most job applications were sent out	-
DataCounty	Country in which data was collected	-
IneqIndex	Gender inequality index for nation of data collection	-
HumanDev	Human development index (retrieved from United Nations Human Development Report)	-
EduIndex	Education index (retrieved from United Nations Human Development Report)	-
GDP	GDP per capita (retrieved from The World Bank)	-
USADist	WEIRD index assessing cultural distance from the United States (Muthukrishna et al., 2020)	-
WithinSubj	Was gender manipulated within-subject (firms received female <i>and</i> male applications) or between-subjects (firms received female <i>or</i> male applications)?	1 = yes (within-subject); 0 = no (between-subject, other)
PairedRealAppl	Real applicants were matched to the best of the researchers' abilities (as opposed to random assignment of fictitious applicant attributes)	1 = yes; 0 = no
Method	Job application method	E.g., resume; other (e.g., in person)
Job	Description of job(s) that was or were applied to	-
JobStereotype	Is the job female-typed, relatively gender-balanced, or male-typed? (as rated by four human coders)	0 = female-typed; 1 = male-typed; 2 = gender balanced
NonFemaleTyped	Dummy variable for job stereotype (as rated by four human coders)	0 = female-typed job; 1 = other
JobStereotype60	Is the job stereotypically male-typed, neutral-typed, or stereotypically female-typed? (objective country-level data; 60% threshold)	0 = female-typed; 1 = male-typed; 2 = gender balanced
NonFemaleTyped60	Dummy variable for non-female-typed jobs (objective country-level data; 60% threshold)	0 = female-typed job; 1 = other

JobStereotype65	Is the job stereotypically male-typed, neutral-typed, or stereotypically female-typed? (objective country-level data; 65% threshold)	0 = female-typed; 1 = male-typed; 2 = gender balanced
NonFemaleTyped65	Dummy variable for non-female-typed jobs (objective country-level data; 65% threshold)	0 = female-typed job; 1 = other
JobStereotype70	Is the job stereotypically male-typed, neutral-typed, or stereotypically female-typed? (objective country-level data; 70% threshold)	0 = female-typed; 1 = male-typed; 2 = gender balanced
NonFemaleTyped70	Dummy variable for non-female-typed jobs (objective country-level data; 70% threshold)	0 = female-typed job; 1 = other
Physical	Does the job require physical strength? (as rated by 4 human coders)	0 = no; 1 = yes
Nurturance	Does the job require nurturance? (as rated by 4 human coders)	0 = no; 1 = yes
ApplEdu	Applicant education (if reported)	-
DV	Operationalization of dependent measure	E.g., call back, interview, job offer
Moderators	Were additional factors (beyond gender) manipulated?	1 = yes; 0 = no
Condition	Condition name if additional factors manipulated other than gender	-
FemaleProp	Proportion of female applicants that succeeded in progressing to the next stage of selection	-
FemaleN	Total number of female applications	-
MaleProp	Proportion of male applicants that succeeded	-
MaleN	Total number of male applications	-
FemaleSucc	Number of female applicants that succeeded	-
FemaleFail	Number of female applicants that did not succeed	-
MaleSucc	Number of male applicants that succeeded	-
MaleFail	Number of male applicants that did not succeed	-
SampleSize	Total sample size (female + male)	-

S4. Study 1: Summary of Changes Made in Response to Red Team Feedback

A group of independent experts (the “red team”) proposed improved approaches to collecting, coding, and analyzing the data throughout the project. Table S3 below provides an overview of the most notable changes we made in response to the feedback by the red team. 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/?view_only=83582fed2db744699361833e6118c7f5).

Table S3. Summary of changes made in response to red team feedback.

Phase	Red Team Feedback	Changes by Authors
Research Question	The red team suggested that the research question could be more precise and preregistered in advance.	We put together a detailed, systematic preregistration following the PROSPERO framework . PROSPERO is an international prospective register of systematic reviews aimed at facilitating rigorous health-related human and animal studies and is considered the gold standard for preregistrations of meta-analyses. Our preregistration is available on Open Science Framework. Among other things, the preregistration provides a detailed overview of the research question and the competing theories being tested.
Article Search	The red team suggested that the search string was too narrow and should be tailored to each database.	We revised the search string to capture combinations of keywords related to gender (e.g., <i>gender</i> , <i>sex*</i> , <i>female*</i> , <i>woman</i> , <i>women</i>), discrimination (e.g., <i>bias</i> , <i>stereotyp*</i> , <i>discriminat*</i> , <i>prejudice*</i>), and field experimental methodology (e.g., <i>audit stud*</i> , <i>field experiment*</i> , <i>randomized trial*</i>). The search term was pretested before preregistration to ensure an optimal balance between effectiveness and efficiency. We also preregistered separate search terms for each database .
	The red team recommended additional databases that should be searched and to be more precise in reporting the exact databases and search parameters.	We added Web of Science Core Collection and Google Scholar to our search. We also indicate all databases that were searched via EBSCO (e.g., PsycArticles) as well as the respective search parameters (e.g., document type) to enhance replicability.
	The red team advised us to use a more thorough citation tracking approach.	We conducted a preregistered, systematic backward citation tracking approach in which we reviewed the references of key empirical papers, reviews, and meta-analysis related to the topic. We also conducted forward citation tracking in which we reviewed all forward citations of the five most highly cited studies in our sample.
	The red team suggested that our initial assumption that there was no substantial file drawer for audit studies was unrealistic and that we should conduct a thorough search for unpublished work .	Our revised search approach included several steps to identify unpublished studies. First, we issued public calls on listservs/ community boards and on social media. Second, we contacted all corresponding authors of the published studies included in our final sample. Finally, we searched for unpublished dissertations in dissertation databases .
	The red team proposed a more systematic study screening and selection process .	We devised systematic, preregistered criteria for study screening and selection . For example, we specified what studies were eligible for inclusion and how/by whom screening decisions would be made. We also followed the red team’s suggestion to use a collaborative online screening platform (i.e., Rayyan).

	The red team identified a typo in the search string for Web of Science and the EBSCO databases in the preregistration.	We have corrected the search string and searched the databases with the updated string . This led to the addition of two new studies.
Coding Scheme	The red team suggested to develop a more transparent and intuitive coding scheme .	We revised and preregistered the coding scheme according to the red team's suggestions. For example, we now provide a detailed description of each variable as well as how the variables were coded.
	The red team had questions regarding the purpose of some of the coded variables .	In the preregistration form, we a priori defined each variable to be extracted as well as which of these variables we planned to use in our analysis and in what capacity.
	The red team highlighted the importance of coding and accounting for certain study characteristics .	We revised or added several important variables to address the red team's suggestions. To account for study location, it was not possible to control for each country or territory separately as the large number of countries would exceed the available degrees of freedom. Thus, we used the Gender Inequality Index (GII) from the United Nations to account for differences in workforce participation by women – the difference across countries most likely to be relevant for our time trend analysis. In addition, we accounted for the type of job that was applied for by coding whether a job was primarily female-typed, male-typed, or balanced/mixed . The last category consisted of jobs that were either gender neutral or a mix of female-typed, gender-balanced, and male-typed jobs without separate effect sizes being reported.
	The red team noticed that in our initial sample of studies, some publications relied on the same data .	In our revised search, we were careful to identify duplicate datasets used across publications. When necessary, we contacted the authors to resolve ambiguities.
	The red team was unsure what effect size metric we were planning to use and how it would be calculated.	We preregistered that we would rely on the log-odds ratio in our analysis, which is calculated based on the raw counts of successes/failures by female vs. male applicants. We also justified why this approach is superior to alternative metrics (e.g., difference of proportions).
Analyses	The red team suggested to only report on multilevel analyses and not report univariate analyses as a multilevel approach is most appropriate.	We now exclusively report multilevel analyses in the main report , which is the most appropriate approach. However, for transparency reasons and because our preregistration did not anticipate a hierarchical data structure, we report univariate analyses in this Supplement .
	The red team identified a duplication in the numbering of studies in the dataset.	We have corrected this and repeated all analyses accordingly.
	The red team suggested that publication bias analyses should also be conducted with the unpublished studies included.	We also applied the publication bias methods to all studies (i.e., including unpublished studies) .
	The red team suggested to extend the PET-PEESE method to multilevel meta-analysis and not apply this method in the context of a univariate meta-analysis.	The PET-PEESE method was extended to a multilevel meta-analysis .
Manuscript & Supplement	The red team recommended to include a “risk of bias” assessment .	We now highlight potential issues with audit studies in general and also highlight the limitations of the studies included in our sample.
	The red team suggested that our discussion could consider the issue of intersectionality .	In the discussion, we now explicitly highlight the possibility that gender and race are not separate dimensions of discrimination, and that workers with multiple marginalized identities (e.g., Black women) can experience unique forms of discrimination that intersects these two identities.

	The red team suggested that we consider how people may respond to this upward trend in women hiring with backlash .	We now discuss the possibility that organizations that experience an upward trend in hiring women may need to brace themselves for backlash against this increased diversity among members of historically privileged groups.
	The red team suggested several passages in which comprehension/interpretation could be better or additional elaboration was needed.	We addressed all comments regarding comprehension and interpretation directly in the respective passages of the manuscript.

S5. Study 1: Robustness Test Using Univariate Random-Effects Meta-Analyses

During data collection, we decided to deviate from the preregistered analyses and take the nesting of effect sizes within studies into account by conducting multilevel meta-analyses. However, in Table S4 we also report the results of the preregistered univariate random-effects meta-analysis. The restricted maximum likelihood estimator (Raudenbush, 2009) was used in the univariate random-effects meta-analyses for estimating the between-study variance in true effect size and the Q -profile method (Viechtbauer, 2007) was used for computing the corresponding confidence interval. The Hartung-Knapp adjustment (Hartung & Knapp, 2001; Sidik & Jonkman, 2002) was used for testing hypotheses and computing confidence intervals of the fixed effects in the univariate random-effects meta-analyses to take uncertainty in the estimated within-study and between-study variances into account.

Table S4. Robustness Test Using Univariate Random-Effects Meta-Analyses. For each model and variable, the parameter estimate, standard error (), 95% confidence interval (CI) [], *t*-value, and corresponding two-tailed *p*-value of the univariate random-effects meta-analyses are displayed. The between-study variance was estimated using the restricted maximum likelihood estimator. The Knapp-Hartung adjustment was applied for testing hypotheses and computing 95% CIs.

	Average Gender Discrimination		Moderation by Gender Typicality of Job		Discrimination Over Time	
	Model 1a	Model 1b	Model 2a	Model 2b	Model 3a	Model 3b
Predictors						
Intercept	-.101 (.025) [-.151;-.051] <i>t</i> (243) = -4.003 <i>p</i> < .001	.031 (.090) [-.208;.146] <i>t</i> (240) = -.343 <i>p</i> = .732	-.288 (.046) [-.378;-.199] <i>t</i> (242) = -6.329 <i>p</i> < .0001	-.224 (.095) [-.412;.036] <i>t</i> (239) = -2.349 <i>p</i> = .020	.285 (.124) [.040;.529] <i>t</i> (169) = 2.299 <i>p</i> = .023	.329 (.134) [.064;.594] <i>t</i> (166) = 2.451 <i>p</i> = .015
Non-female-typed job			.249 (.054) [.154;.365] <i>t</i> (242) = 4.84 <i>p</i> < .0001	.260 (.055) [.152;.367] <i>t</i> (239) = 4.763 <i>p</i> < .0001		
Application year					-.009 (.003) [-.016;-.002] <i>t</i> (169) = -2.595 <i>p</i> = .010	-.014 (.005) [-.023;-.005] <i>t</i> (166) = -3.185 <i>p</i> = .002
Controls						
Inequality index		-.089 (.251) [-.583;.405] <i>t</i> (240) = -.354 <i>p</i> = .723		-.215 (.241) [-.690;.260] <i>t</i> (239) = -.891 <i>p</i> = .374		-.288 (.255) [-.790;.215] <i>t</i> (166) = -1.131 <i>p</i> = .260
Study design complexity ^a		-.020 (.083) [-.184;.144] <i>t</i> (240) = -.244 <i>p</i> = .807		-.010 (.080) [-.167;.148] <i>t</i> (239) = -.124 <i>p</i> = .902		.198 (.113) [-.026;.421] <i>t</i> (166) = 1.745 <i>p</i> = .083
Proportion female authors		-.100 (.078) [-.253;.053] <i>t</i> (240) = -1.288 <i>p</i> = .199		-.063 (.074) [-.209;.083] <i>t</i> (239) = -.854 <i>p</i> = .394		.034 (.080) [-.125;.193] <i>t</i> (166) = .423 <i>p</i> = .673
$\hat{\tau}^2$ [95% CI] ^b	.097 [.077;.135]	.097 [.077;.137]	.084 [.066;.119]	.085 [.066;.120]	.072 [.051;.104]	.072 [.051;.103]
I^2 [95% CI] ^b	.824 [.788;.868]	.822 [.786;.867]	.802 [.761;.852]	.800 [.759;.851]	.807 [.748;.860]	.802 [.740;.853]
<i>Q</i> -statistic; <i>p</i> -value	<i>Q</i> (243) = 1129.24 <i>p</i> < .0001	<i>Q</i> (240) = 1202.96 <i>p</i> < .0001	<i>Q</i> (242) = 1141.19 <i>p</i> < .0001	<i>Q</i> (239) = 1123.13 <i>p</i> < .0001	<i>Q</i> (169) = 819.04 <i>p</i> < .0001	<i>Q</i> (166) = 804.91 <i>p</i> < .0001

^a No moderators is the reference category. ^b The 95% CI was computed using the *Q*-profile method.

S6. Study 1: Sensitivity Analyses of Alternative Gender Typicality Thresholds

In addition to the commonly used approach in which 65% is used as threshold to categorize a job as female- or male-typed, we also explored alternative thresholds below (60%) and above (70%) the one reported in the main text to examine the robustness of our findings.

Table S5. Sensitivity Analyses of Alternative Gender Typicality Thresholds. For each model and variable, the parameter estimate, standard error (), 95% confidence interval (CI) [], *t*-value, and corresponding two-tailed *p*-value of the univariate random-effects meta-analyses are displayed. The between-study variance was estimated using the restricted maximum likelihood estimator. The Knapp-Hartung adjustment was applied for testing hypotheses and computing 95% CIs.

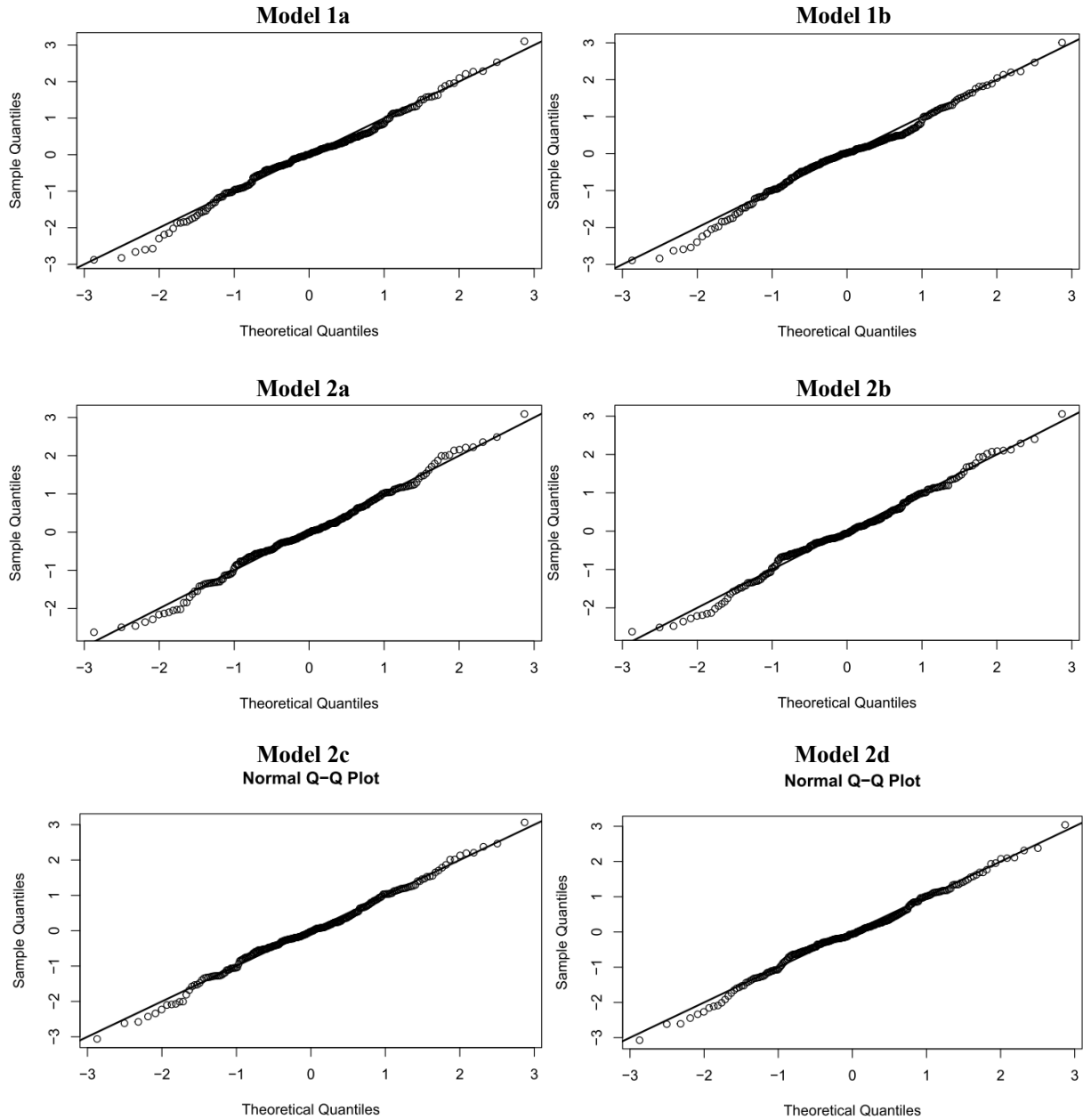
	Objective Coding Approach (60% cutoff, exploratory)				Objective Data Approach (70% cutoff, exploratory)			
	Gender Typicality of Job		Time Trend		Gender Typicality of Job		Time Trend	
	Model 2a	Model 2b	Model 3a	Model 3b	Model 2c	Model 2d	Model 3c	Model 3d
Intercept	-.246 (.046) [-.335;-.157] <i>z</i> = -5.393 <i>p</i> < .001	-.179 (.117) [-.408;.049] <i>z</i> = -1.539 <i>p</i> = .124	.281 (.142) [.003;.559] <i>z</i> = 1.984 <i>p</i> = .047	.326 (.159) [.016;.637] <i>z</i> = 2.059 <i>p</i> = .040	-.239 (.063) [-.363;-.115] <i>z</i> = -3.769 <i>p</i> < .001	-.158 (.131) [-.414;.099] <i>z</i> = -1.207 <i>p</i> = .228	.266 (.158) [-.044;.576] <i>z</i> = 1.683 <i>p</i> = .035	.329 (.174) [-.012;.670] <i>z</i> = 1.891 <i>p</i> = .059
Non-female-typed job	.223 (.051) [.124;.322] <i>z</i> = 4.401 <i>p</i> < .001	.221 (.051) [.152;.358] <i>z</i> = 4.319 <i>p</i> < .001			.176 (.065) [.048;.303] <i>z</i> = 2.705 <i>p</i> = .007	.176 (.065) [.048;.304] <i>z</i> = 2.693 <i>p</i> = .007		
Application year			-.009 (.004) [-.016;-.001] <i>z</i> = -2.187 <i>p</i> = .029	-.012 (.005) [-.022;-.002] <i>z</i> = -2.257 <i>p</i> = .024			-.009 (.004) [-.018;-.001] <i>z</i> = -2.108 <i>p</i> = .035	-.013 (.006) [-.024;-.002] <i>z</i> = -2.312 <i>p</i> = .021
Inequality index		-.181 (.283) [-.735;.373] <i>z</i> = -.640 <i>p</i> = .522		-.289 (.294) [-.865;.287] <i>z</i> = -.984 <i>p</i> = .325		-.197 (.301) [-.786;.393] <i>z</i> = -.653 <i>p</i> = .514		-.337 (.314) [-.951;.278] <i>z</i> = -1.075 <i>p</i> = .283
Study design complexity ^a		-.007 (.102) [-.207;.192] <i>z</i> = -.071 <i>p</i> = .943		.118 (.137) [-.152;.387] <i>z</i> = .856 <i>p</i> = .392		-.017 (.110) [-.233;.200] <i>z</i> = -.151 <i>p</i> = .880		.150 (.143) [-.130;.430] <i>z</i> = 1.048 <i>p</i> = .295
Proportion female authors		-.075 (.083) [-.237;.088] <i>z</i> = -.901 <i>p</i> = .368		.029 (.094) [-.156;.214] <i>z</i> = .308 <i>p</i> = .758		-.087 (.088) [-.259;.084] <i>z</i> = -.998 <i>p</i> = .318		-.055 (.095) [-.241;.132] <i>z</i> = -.573 <i>p</i> = .567
$\hat{\sigma}_1^2$ [95% CI]	.02 [.00;.05]	.02 [.00;.05]	.01 [.00;.05]	.01 [.00;.05]	.02 [.00;.06]	.03 [.00;.06]	.02 [.00;.06]	.03 [.00;.06]
$\hat{\sigma}_2^2$ [95% CI]	.07 [.05;.10]	.07 [.05;.10]	.08 [.05;.12]	.08 [.05;.12]	.08 [.05;.11]	.07 [.05;.11]	.08 [.05;.11]	.08 [.05;.11]
<i>I</i> ² -statistic	0.812	0.813	0.832	0.830	0.826	0.826	0.841	0.839
<i>Q</i> -statistic, <i>p</i> -value	1149.87 <i>p</i> < .0001	1130.25 <i>p</i> < .0001	833.63 <i>p</i> < .0001	824.63 <i>p</i> < .0001	1209.15 <i>p</i> < .0001	1186.57 <i>p</i> < .0001	1053.61 <i>p</i> < .0001	1038.13 <i>p</i> < .0001

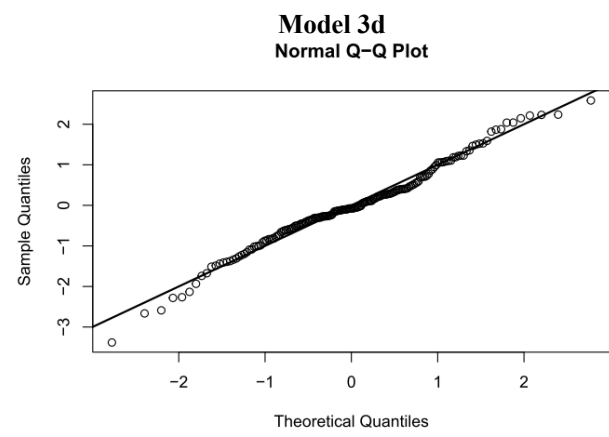
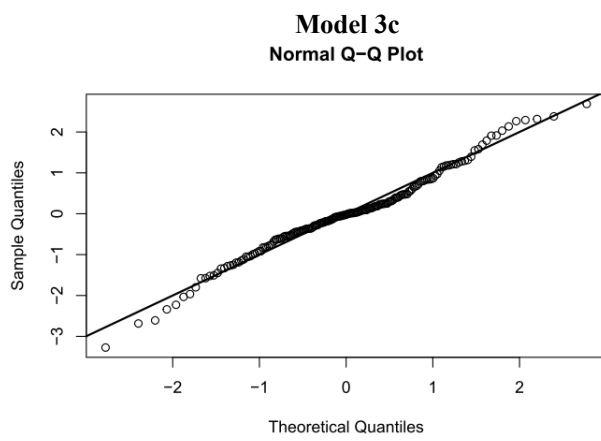
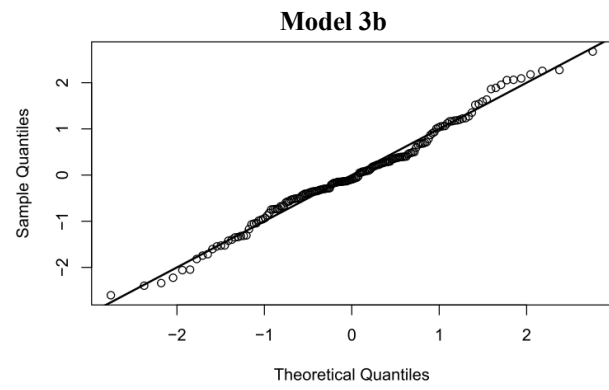
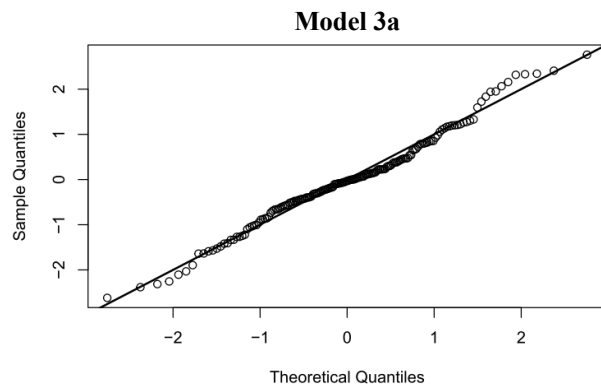
^a No moderators is the reference category.

S7. Study 1: Assessment of Normality Assumptions Multilevel Meta-Analysis Model

We verified whether the fitted multilevel meta-analysis models was an appropriate approximation of the data by creating Q-Q plots (Wang & Bushman, 1998). We created Q-Q plots with the theoretical quantiles of the normal distribution on the x -axis and the standardized residuals on the y -axis (see Fig. S1), none of which indicate clear violations.

Fig. S1. Q-Q plots. The Q-Q plots indicate whether the fitted multilevel meta-analysis models was an appropriate approximation of the data. One Q-Q plot is presented for each statistical model (see Table 2 in the main manuscript for an overview).





S8. Study 1: Influential Effect Sizes

We also checked for influential effect sizes in each fitted multilevel meta-analysis model. We computed externally standardized residuals (Viechtbauer & Cheung, 2010), examined the weight that was assigned to each effect size in the meta-analyses, and conducted leave-one-out analyses where the multilevel meta-analysis was repeatedly fitted to the data and each time a single study was discarded. Fig. S2 and S3 represent the distribution of the estimated effects of the time variable and corresponding p -values indicating whether the average effect size is different from 0 when each observation was individually excluded from the data. We concluded that there were no influential effect sizes in the meta-analyses for Models 1 and 2 (see Table 2 in main text), because the most extreme externally standardized residual was 3.23 and the largest weight of a study was 0.7%.

There was an influential effect size for Model 3 (see Table 2 in main text), because the statistically significant negative relationship of time was no longer statistically significant if the effect size with UniqueID 14A was entirely discarded from the meta-analysis (Model 3a: -0.005 , $z = -1.26$, $p = .213$; Model 3b: -0.004 , $z = -.70$, $p = .481$). The externally standardized residual of this study was 2.544 in both Model 3a and Model 3b. This is the effect size based on the data that were collected in 1978 (odds ratio of 2.538) in the top left corner of Fig. 2 in the main text.

In Fig 6 (see main text), there are also two very small negative effect sizes, which come from observations UniqueID 16D and 16E. These effect sizes were highly negative and the corresponding sampling variance was small, because none of the males had success when applying for a job. These effect sizes only had a minor impact on the meta-analytic results due to their high sampling variances.

Overall, our analyses were not highly influenced by a single study, with the exception of one large study conducted in 1978, which is not surprising given the number of audit studies before 1990 was small (for a similar conclusion for race studies, see Quillian, Pager, Hexel, & Midtbøen, 2017). See the General Discussion of the main manuscript for further elaboration on this issue.

Fig. S2. Distribution of estimated effects for leave-one-out analyses.

The plots indicate the frequency of estimated effects for each preregistered statistical model (see Table 2 in the main text for an overview) when each observation in the dataset is excluded one by one.

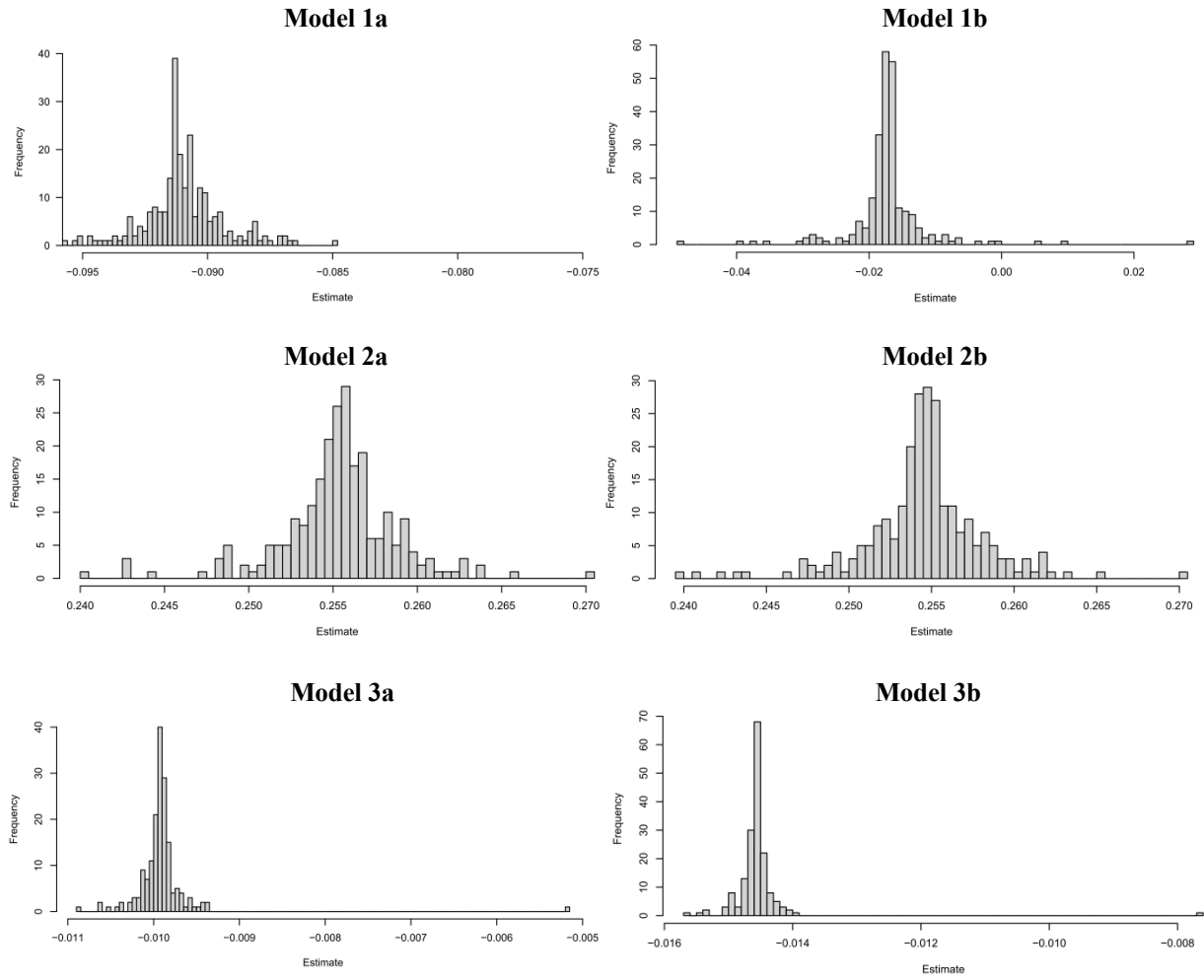
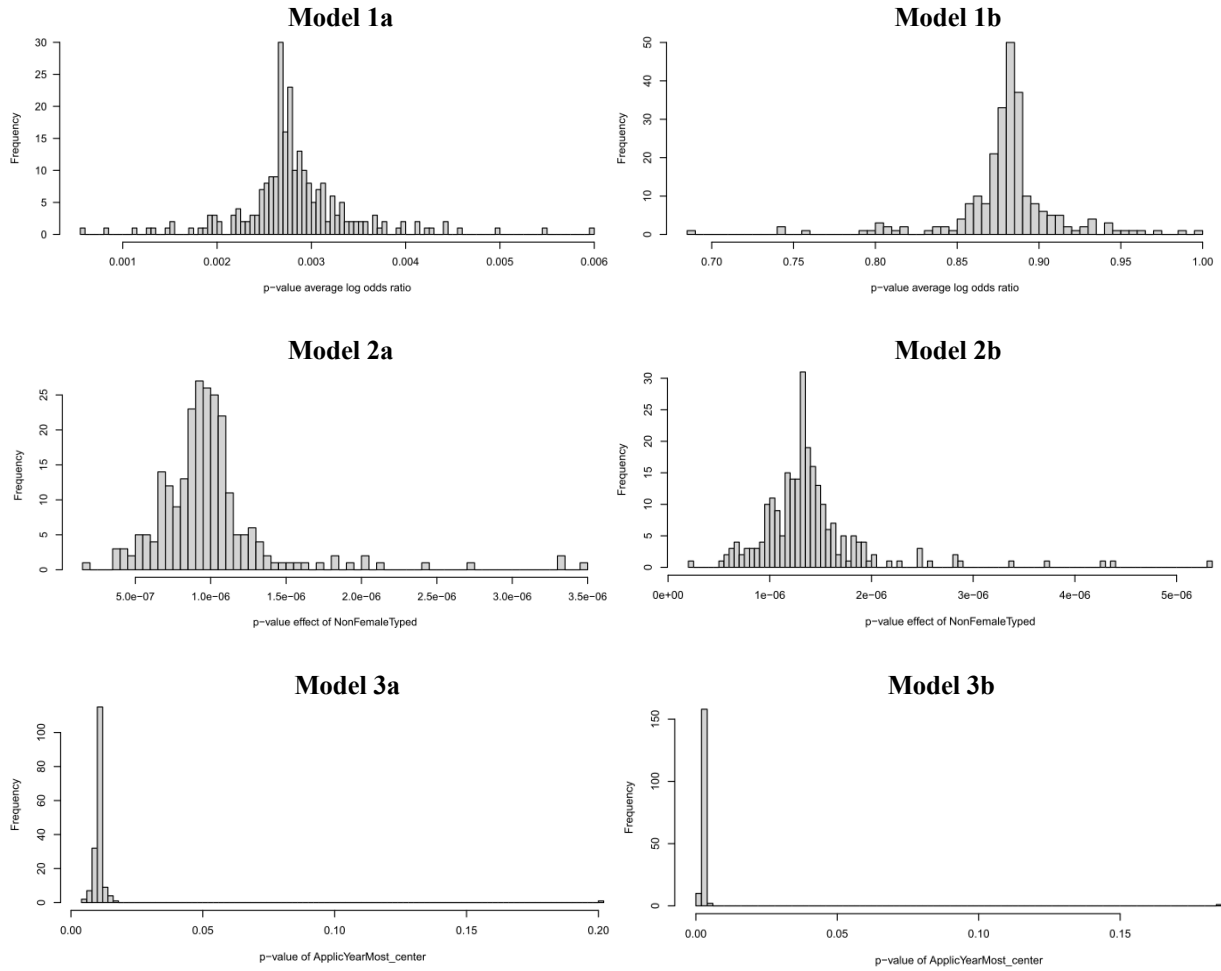


Fig. S3. Distribution of p -values for leave-one-out analyses.

The plots indicate the frequency of p -values for each preregistered statistical model (see Table 2 in the main text for an overview) when each observation in the dataset is excluded one by one.



S9. Study 1: Red Team Member Profiles

To implement the red team approach, we collaborated with two “coordinators” who in turn helped us recruit five “red team reviewers” who are experts in the domain of gender, field experimental methods, meta-analytic methods, and/or statistics. One gender expert is a qualitative researcher. Below, we provide short biographies of the two coordinators and five red team members (four female, one male) and highlight their relevant qualifications.

Red Team Coordinator 1. This scholar is a tenured professor in human-technology interaction at a research university in the Netherlands. They are an expert on meta-science, research methods, and applied statistics. They have published numerous articles in general science and psychology journals.

Red Team Coordinator 2. This scholar is a postdoctoral researcher in the human-technology interaction group at a research university in the Netherlands. They are an expert in meta science and study the forces that shape how scientists do their work and make inferences from the populations that they study. They have published numerous articles in general science and psychology journals.

Red Team Reviewer 1. This scholar is an expert in gender research and field experimental methods and is a full professor at a research university the United States. They have published numerous articles related to workplace discrimination and the challenges of stigmatized groups in general science, management, and psychology journals.

Red Team Reviewer 2. This scholar is an expert in diversity and gender research and is currently an advanced doctoral student in social psychology at a research university in the United States. They have published their research in psychology journals.

Red Team Reviewer 3. This scholar is an expert in gender and human resources research. They are a postdoctoral researcher in political science at a research university in Denmark. They have published several articles in organizational behavior journals.

Red Team Reviewer 4. This scholar is an expert in statistics and meta-analytic methods. They hold a doctorate degree in education and are a research scientist in meta-science at a research university in Slovakia. They have published numerous articles in organizational behavior and social psychology journals.

Red Team Reviewer 5. This scholar is an expert in bibliographic search. They are a senior librarian at a research university in the Netherlands. They have co-authored several research publications in psychology and health-related journals.

S10. Study 2: Pre-Registered Analysis Plan for Forecasting Survey

The original preregistration form is available on the Open Science Framework (https://osf.io/fnxu4/?view_only=b096e2b2f3094bb3bcb082cf7e366f28).

In this project, we examine the extent to which scientists and laypeople can predict the results of a meta-analysis of field audits of gender discrimination conducted between 1976 and 2020. We will recruit two different samples that will fill out a forecasting survey. The first sample is a sample of scientists and the second sample is a nationally representative sample of the US from Prolific Academic (we refer to this sample as “laypeople” below). We describe each of these two samples/data collections in turn below.

For the scientists sample we aim to recruit as many forecasters as possible, setting the goal of achieving an N of at least 50 forecasters. If we fail to reach a sample size of at least $N=50$ scientists that submit a forecast to all the forecasting questions we will not carry out any of the analyses below involving the scientists sample. The first round of data collection for the scientists sample, running in November 2021 will involve up to 14 PhD students and one Research Assistant from different academic areas (e.g., strategy, finance, operations, organizational behavior) in a core doctoral course. We do this round first to ensure that there are no problems with the data collection. If we do not encounter any problems in this first wave of data collection that lead to changes to the survey, these observations will be included in the analyses described below in the scientists sample. If we make any changes to the survey, these first observations will be excluded. For the main data collection of the scientists sample, we plan to advertise the survey via social media (e.g., Twitter, Facebook) and professional listservs (e.g., American Psychological Society) and email social science professors directly. The forecasting survey for the scientists sample will include a question about whether or not the individual is in academia (yes/no) and only individuals answering yes to this question will be included in the scientists sample and the analyses below. We plan to keep the scientists sample data collection open for 4 weeks, and participants will have 28 days to finish the survey. Reminders will be sent out 14 and 7 days prior to the expiration of the survey. We will not analyze the data until all of it has been collected. Only participants who submit forecasts to all the forecasting questions (and answer yes to the question of whether they are in academia; see above) will be included in the analyses for the scientists sample. Participants in the scientists sample who fully complete the

survey and answer yes to the question about if they are in academia will be offered a consortium authorship credit (“Gender Audits Forecasting Collaboration”) in the main author string, with full names and affiliations listed in an appendix to the manuscript.

For our sample of laypeople, we will recruit a nationally representative sample of the US from Prolific Academic (an online survey platform) of $n=500$ that will receive the forecasting survey. Participants will receive a small cash payment of \$2.50 USD for completing the forecasting survey. Only participants in this sample who submit forecasts to all the forecasting questions will be included in the analyses for the laypeople sample.

The forecasters’ task is to predict gender bias in hiring decisions from studies over the past 44 years. The results to be predicted come from an analysis of all available experimental audit studies ($N=87$) from 1976 to 2020 in which identical (or very similar) applications were submitted by either a female candidate or a male candidate and callback were recorded. Gender bias is measured from whether applications that were submitted by women were more or less likely to receive callbacks than those submitted by men. The jobs have been coded as stereotypically female, stereotypically male, or gender neutral.

The meta-analysis will measure discrimination in terms of the log odds ratio of receiving callbacks for men versus women (with an (log) odds ratio above 1 (0) implying that men are favored and women are discriminated against and a log odds ratio below 1 (0) implying that women are favored and men discriminated against). But as we think it would be too demanding to ask forecasters to predict the log odds ratio or the odds ratio, we will ask forecasters to predict the % callback between 0 and 100% for men and for women. These responses will then be converted to the log odds ratio in all the analyses below with a log odds ratio above 0 implying a prediction that men are favored and a log odds ratio below 0 implying a prediction that women are favored (i.e. in all the analysis below the response of an individual forecaster is converted to the implied log odds ratio; e.g. if a forecaster believe that men will receive 15% callbacks and women will receive 10% callbacks the implied log odds ratio is: $\ln[(15/85)/(10/90)]=\ln(1.59)=0.46$).

Forecasters will be asked to predict how much gender bias there will be for 1) female-typed jobs and 2) male-typed jobs plus neutral-typed jobs. For each type of job, forecasters will be asked to predict the results of those studies from the following four time periods; 1976-1986, 1987-1997, 1998-2008 and 2009-2020. For each forecasting question they will be asked to

predict the % callback rate for men and for women respectively (which is then converted to the predicted log odds ratio for the analysis as noted above).

In addition to the 8 forecasting questions described above, forecasters will also be asked about their predictions for female-typed jobs and male-typed jobs plus neutral-typed jobs for the entire time period (these two forecasting questions are referred to as “entire period female-typed job forecast” and “entire period male-typed plus neutral-typed job forecast” below). Finally, forecasters will also be asked to predict the result for the entire sample of studies included in the meta-analysis, pooling the female-typed jobs and the male-typed plus neutral-typed jobs into an overall estimate for the entire time period (this forecast is referred to as “pooled discrimination forecast” below). Forecasters will thus respond to 11 forecasting questions in total. Forecasters in both samples will also be asked about political orientation, system justification and gender system justification, and forecasters in the scientists sample will be asked whether or not they are in academia (see above) and whether they consider themselves to be a gender researcher.

In the hypothesis tests described below based on the data from the forecasting survey, we use both the more conservative significance threshold of $p < .005$ proposed by Benjamin, Berger, Johannesson, Nosek, Wagenmakers, Berk et al. (2018) and the traditional threshold for statistical significance of $p < .05$. Readers can make their own decision regarding which threshold they wish to apply. All the tests in this pre-analysis plan will report two-tailed p-values.

Our tests below are divided into primary hypotheses, secondary hypotheses, and exploratory analyses. As noted above all statistical tests below are carried out in terms of predicted log odds ratios and when forecasts are compared to estimated results from the meta-analysis the estimated results are also in terms of log odds ratios (with a positive log odds ratio implying that men are favored and a negative log odds ratio implying that women are favored).

All the hypotheses and exploratory analyses below will be estimated separately for the two samples of forecasters: the sample of scientists and the sample of laypeople. The exception to this is secondary hypothesis 4 and exploratory analysis 8 that is about comparing results between the two samples of forecasters (scientists and laypeople); and exploratory analysis 4 that compares two groups within the scientist sample.

Primary Hypotheses

Primary Hypothesis 1a: Forecasters believe that men experience more positive job application outcomes than women. This is tested by testing if the average forecast on the “pooled discrimination forecast” question differs significantly from 0 in a one-sample t-test.

Primary Hypothesis 1b: Forecasters under/overestimate the meta-analytic gender discrimination. This is tested by testing if the average forecast on the “pooled discrimination forecast” question differs significantly from the estimated gender discrimination in the meta-analysis (the meta-analysis estimate for all jobs (the female-typed jobs, the neutral-typed jobs and the male-typed jobs) for the entire period studied). A z-test will be used. We have no hypothesized direction of this test.

Primary Hypothesis 2a: Forecasters believe that discrimination of women relative to men is larger in male-typed plus neutral-typed jobs than in female-typed jobs. This is tested by testing if the average forecast on the “entire period male-typed plus neutral-typed job forecast” question differs significantly from the “entire period female-typed job forecast” question. A paired t-test will be used (as all forecasters answer both forecasting questions the data is paired for this test).

Primary Hypothesis 2b: Forecasters under/overestimate the meta-analytic difference in gender discrimination between male-typed plus neutral-typed jobs and female-typed jobs. This is a “difference-in-difference” test comparing the forecasted difference in discrimination between the “entire period male-typed plus neutral-typed job forecast” question and the “entire period female-typed job forecast” question and the estimated difference in discrimination between “male-typed plus neutral-typed jobs” and “female-typed jobs”.

To estimate the forecasted difference, we first estimate the difference between the “entire period male-typed plus neutral-typed job forecast” question and the “entire period female-typed job forecast” question for each forecaster and then we estimate the mean and standard error of this difference (this is the same as the effect size and standard error in the paired t-test in primary hypothesis 2a above).

The second difference is the estimated difference between “male-typed plus neutral-typed jobs” and “female-typed jobs” in the meta-analysis for the entire time period. This difference is estimated in a z-test based on the meta-analytic estimate for “male-typed plus neutral-typed jobs” and its standard error and the meta-analytic estimate for “female-typed jobs” and its standard error.

The “difference-in-difference” test is carried out as a z-test based on the effect (difference) and the standard error of the two differences defined above. We have no hypothesized direction of this test.

Primary Hypothesis 3a: Forecasters believe that discrimination against women relative to men has decreased over time for male-typed plus neutral-typed jobs. This is tested by testing if the average forecast on the male-typed plus neutral-typed job forecast question differs between the first period and the last period. A paired t-test is used (as all forecasters answer both forecasting questions the data is paired for this test).

Primary Hypothesis 3b: Forecasters under/overestimate the meta-analytic reduction in gender discrimination over time for male-typed plus neutral-typed jobs. This is a “difference-in-difference” test comparing the forecasted difference in discrimination between the first and the last period for the male-typed plus neutral-typed job forecast question and the estimated meta-analytic difference in discrimination between the first and the last period for male-typed plus neutral-typed jobs.

To estimate the forecasted difference, we first estimate the difference between the male-typed plus neutral-typed jobs question in the first period and the male-typed plus neutral-typed jobs question in the last period for each forecaster and then we estimate the mean and standard error of this difference (this is the same as the effect size and standard error in the paired t-test in primary hypothesis 3a above).

The second difference is estimated as the difference in the meta-analytic effect size for male-typed plus neutral-typed jobs between the first period and the last period. This difference is estimated in a z-test based on the meta-analytic estimate and its standard error in these two periods.

The “difference-in-difference” test is carried out as a z-test based on the effect (difference) and the standard error of the two differences defined above. We have no hypothesized direction of this test.

Secondary Hypotheses

Secondary hypotheses 1-2: In these tests we carry out the same tests as in primary hypothesis 1a and 1b; but separately for the female-typed jobs and the male-typed plus neutral typed jobs.

Secondary Hypothesis 1a: Forecasters believe that men experience more positive job application outcomes than women for male-typed plus neutral typed jobs. This is tested by testing if the average forecast on the “entire period male-typed plus neutral-typed job forecast” question differs significantly from 0 in a one-sample t-test.

Secondary Hypothesis 1b: Forecasters under/overestimate the meta-analytic gender discrimination for male-typed plus neutral typed jobs. This is tested by testing if the average forecast on the “entire period male-typed plus neutral-typed job forecast” question differs significantly from the estimated gender discrimination in the meta-analysis (the meta-analysis estimate for the male-typed plus neutral-typed jobs for the entire period studied). A z-test will be used. We have no hypothesized direction of this test.

Secondary Hypothesis 2a: Forecasters believe that women experience more positive job application outcomes than men for female-typed jobs. This is tested by testing if the average forecast on the “entire period female-typed job forecast” question differs significantly from 0 in a one-sample t-test.

Secondary Hypothesis 2b: Forecasters under/overestimate the meta-analytic gender discrimination for female-typed jobs. This is tested by testing if the average forecast on the “entire period female-typed job forecast” question differs significantly from the estimated gender discrimination in the meta-analysis (the meta-analysis estimate for the female-typed jobs for the entire period studied). A z-test will be used. We have no hypothesized direction of this test.

Secondary Hypothesis 3: There is a positive association between the predictions (beliefs) of forecasters and the meta-analytic results. In secondary hypotheses 3 we test if there is a statistically significant positive association between the predicted results and the estimated meta-analytic results. We include the four time period predictions of female-typed jobs and the four time period predictions of male-typed plus neutral-typed jobs in this analysis (i.e. eight observations per forecaster). We will 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 (there will be eight observations per forecaster in the OLS regression and the total number of observations in the OLS regression will equal the number of forecasters times eight). We will include individual fixed effects in the OLS regression and we will cluster 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). The test of this

hypothesis is carried out as a t-test on the regression coefficient of the forecast variable and it will test if forecasters can to some extent predict the variation in the discrimination effect between the four time periods and between the female-typed and male-typed plus neutral typed jobs.

Secondary Hypothesis 4: The accuracy of predictions differs between scientists and laypeople. In our fourth secondary hypothesis we will test if the accuracy of the predictions differ between scientists and laypeople. For each survey-taker, the accuracy of each forecasting question is defined as the squared prediction error (Brier score), where the prediction error is defined as the difference between the prediction and the observed estimate in the meta-analysis.

In this test we first estimate the mean squared prediction error (Brier score) of each forecaster for the 11 predictions (i.e. for each forecaster we estimate the squared prediction error for the 11 predictions made by that forecaster and then we estimate the mean of these 11 squared prediction errors for each forecaster). We then test if the mean squared prediction error differs between scientists and laypeople using an independent samples t-test. We have no hypothesized direction of this test.

Secondary hypothesis 5: The accuracy of predictions differs for the predictions of female-typed jobs and the predictions of male-typed plus neutral-typed jobs. In our fifth secondary hypothesis we will test if the accuracy of the predictions differs between the predictions of female-typed jobs and the predictions of male-typed and neutral-typed jobs.

In this test we first construct two individual level variables. The first of these variables is the mean squared prediction error of each forecaster for the five predictions for the female-typed jobs (i.e. for each forecaster we estimate the squared prediction error for the 5 predictions of female-typed jobs made by that forecaster and then we estimate the mean of these five squared prediction errors for each forecaster). The second of these variables is the mean squared prediction error of each forecaster for the five predictions of the male-typed plus neutral-typed jobs (i.e. for each forecaster we estimate the squared prediction error for the 5 predictions of male-typed plus neutral-typed jobs made by that forecaster and then we estimate the mean of these five squared prediction errors for each forecaster). We then carry out a paired t-test (n =number of forecasters) of these two variables to test if the mean squared prediction error differs for the predictions of female-typed jobs and the predictions of neutral-typed and male-typed jobs. We have no hypothesized direction of this test.

Exploratory Analyses

Exploratory analysis 1: In exploratory analysis 1 we test if the accuracy of predictions varies with political orientation. In this test we first construct two individual level variables. The first is the mean squared prediction error (Brier score) of each forecaster for the 11 predictions (i.e., for each forecaster we estimate the mean squared prediction error for the 11 predictions made by that forecaster and then we estimate the mean of these 11 squared prediction errors for each forecaster). The second variable is a variable measuring political orientation of the forecaster from “very liberal” to “very conservative; this measure is estimated as the average for the forecaster of three 7-point scale survey questions about political orientation with the scale of each attitude question being measured from “very liberal” (-3) to “very conservative” (3) (with 0 as the neutral midpoint). The three political orientation questions are phrased as: “If I had to describe my political views overall, I would say that I am...”; With respect to economic matters, I consider myself...”; With respect to social matters, I consider myself...”.

We then 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 will be estimated with robust standard errors). The test is carried out as a t-test on the regression coefficient of the political orientation variable in the OLS regression. We have no hypothesized direction of this test.

Exploratory analysis 2: In exploratory analysis 2 we test if the accuracy of predictions varies with system justification. In this test we first construct two individual level variables. The first is the mean squared prediction error (Brier score) of each forecaster for the 11 predictions (i.e., for each forecaster we estimate the mean squared prediction error for the 11 predictions made by that forecaster and then we estimate the mean of these 11 squared prediction errors for each forecaster). The second variable is a variable measuring system justification estimated as the average for the forecaster of eight 7-point scale survey questions.

We then estimate an OLS regression with the mean squared prediction error of each forecaster as the dependent variable and the system justification variable as the independent variable (the OLS regression will be estimated with robust standard errors). The test is carried

out as a t-test on the regression coefficient of the system justification variable in the OLS regression. We have no hypothesized direction of this test.

Exploratory analysis 3: In exploratory analysis 3 we test if the accuracy of predictions varies with gender system justification. In this test we first construct two individual level variables. The first is the mean squared prediction error (Brier score) of each forecaster for the 11 predictions (i.e., for each forecaster we estimate the mean squared prediction error for the 11 predictions made by that forecaster and then we estimate the mean of these 11 squared prediction errors for each forecaster). The second variable is a variable measuring gender system justification estimated as the average for the forecaster of eight 7-point scale survey questions.

We then estimate an OLS regression with the mean squared prediction error of each forecaster as the dependent variable and the gender system justification variable as the independent variable (the OLS regression will be estimated with robust standard errors). The test is carried out as a t-test on the regression coefficient of the gender system justification variable in the OLS regression. We have no hypothesized direction of this test.

Exploratory analysis 4: In exploratory analysis 4 we test if the accuracy of predictions differs between scientists with backgrounds studying gender bias and other scientists. This test will only be carried out for the scientist sample of forecasters. In this test we first estimate the mean squared prediction error (Brier score) of each forecaster for the 11 predictions (i.e., for each forecaster we estimate the mean squared prediction error for the 11 predictions made by that forecaster and then we estimate the mean of these 11 squared prediction errors for each forecaster). We then divide the scientist sample of forecasters into scientists with backgrounds studying gender bias and other scientists (based on the survey question: “Do you consider yourself to be a gender researcher? Yes, No”); and test if the mean squared prediction error differs between these two groups using an independent samples t-test. We have no hypothesized direction of this test.

Exploratory analysis 5: In exploratory analysis 5 we will carry out the tests in primary hypothesis 2a and 2b separately for each of the four time periods.

Exploratory analysis 6: In exploratory analysis 6 we will carry out the tests in secondary hypothesis 1a and 1b separately for each of the four time periods.

Exploratory analysis 7: In exploratory analysis 7 we will carry out the tests in secondary hypothesis 2a and 2b separately for each of the four time periods.

Exploratory analysis 8: In exploratory analysis 8 we will compare the mean forecast between the sample of scientists and the sample of laypeople for each of the 11 forecasting questions (with one separate test for each question). An independent sample t-test will be used.

S11. Study 2: Forecasting Survey Items

Study Instructions

A forthcoming research investigation tested for gender biases in hiring decisions. The researchers analyzed all available studies from 1976 to 2020 in which nearly identical applications were submitted to real employers by either a male candidate or a female candidate and callbacks were recorded (e.g., interview invites, job offers). Gender bias was measured by differences in callback rates for applications with male names compared to the same applications with female names.

The researchers summarized the results of 87 such studies from 26 countries. They classified the jobs as "stereotypically female" (e.g., receptionist, dental assistant, and preschool teacher), "stereotypically male" (e.g., truck driver, construction worker, and computer programmer), or "gender neutral" (e.g., either a mix of job types that are gender balanced overall or relatively gender neutral jobs, such as cleaner, accountant, and high-school teacher).

In the present survey, we would like you to predict what the researchers found. In other words, how likely were women to receive callbacks compared to men within the different job domains? We also would like you to estimate these rates within different time periods and overall.

You will be asked to forecast the results for male-typed and neutral-typed jobs together, and female-typed jobs separately. This is because, based on theories of gender stereotyping, the researchers anticipated that there could be different results for female-typed jobs relative to the other categories. By grouping the job categories this way, your predictions will be aligned with the researchers' pre-specified analyses.

The study methodology is available here: [\[link presented here\]](#)

Forecasting Male and Gender Neutral Callbacks

Below, you will estimate callback rates for males and females for **stereotypically male jobs and gender neutral jobs** within four time chunks, starting from 1976 and ending in the present day, as well as overall. Examples of jobs coded as stereotypically male or gender neutral are truck driver, computer programmer, cleaner, or accountant.

In the first column, you will estimate the average percentage of female applicants that received callbacks within each time period, and in the second column, you will estimate the average percentage of male applicants who received callbacks within each time period.

Your responses can range from 0 to 100. If you think women were more likely to receive callbacks (pro-female bias), the percentage should be higher in the first column than the second column. If you think men were more likely to receive callbacks (pro-male bias), the percentage

should be higher in the second column than the first column. If you think there was no gender bias, your percentages in the two columns should match. But note your responses can differ in between time chunks if you believe gender biases have changed over time.

In the final row, you will estimate the overall percentages of callbacks for men and women across all studies from 1976-2020. For these estimates, recall from the redacted methods that most studies occurred in 2005 or later.

As a reminder, **your estimates on this page are for callbacks for jobs that are perceived as stereotypically male or gender neutral.**

	Percentage of women who received callbacks	Percentage of men who received callbacks
1976-1986		
1987-1997		
1998-2008		
2009-2020		
Average over all studies		

Forecasting Female Callbacks

Below, you will estimate callback rates for males and females for **stereotypically female jobs** within four time chunks, starting from 1976 and ending in the present day, as well as overall. Examples of jobs coded as stereotypically female are receptionist, dental assistant, and preschool teacher.

In the first column, you will estimate the average percentage of female applicants that received callbacks within each time period, and in the second column, you will estimate the average percentage of male applicants who received callbacks within each time period.

Your responses can range from 0 to 100. If you think women were more likely to receive callbacks (pro-female bias), the percentage should be higher in the first column than the second column. If you think men were more likely to receive callbacks (pro-male bias), the percentage should be higher in the second column than the first column. If you think there was no gender bias, your percentages in the two columns should match. But note your responses can differ in between time chunks if you believe gender biases have changed over time.

In the final row, you will estimate the overall percentages of callbacks for men and women across all studies from 1976-2020. For these estimates, recall from the redacted methods that most studies occurred in 2005 or later.

As a reminder, **your estimates on this page are for callbacks for jobs that are perceived as stereotypically female.**

	Percentage of women who received callbacks	Percentage of men who received callbacks
1976-1986		
1987-1997		
1998-2008		
2009-2020		
Average over all studies		

Gender System Justification Scale

Please specify the degree to which you agree or disagree with the following statements.
(Response options: 1 = Strongly disagree to 7 = Strongly agree)

1. In general, relations between men and women are fair.
2. The division of labor in families generally operates as it should.
3. Gender roles need to be radically restructured.
4. For women, my country is the best country in the world to live in.
5. Most policies relating to gender and the sexual division of labor serve the greater good.
6. Everyone (male or female) has a fair shot at wealth and happiness.
7. Sexism in society is getting worse every year.
8. Society is set up so that men and women usually get what they deserve.

System Justification Scale

Please specify the degree to which you agree or disagree with the following statements.
(Response options: 1 = Strongly disagree to 7 = Strongly agree)

1. In general, society is fair.
2. In general, the political system in my country operates as it should.
3. Society needs to be radically restructured.
4. My country is the best country in the world to live in.
5. Most policies serve the greater good.
6. Everyone has a fair shot at wealth and happiness.
7. Our society is getting worse every year.
8. Society is set up so that people usually get what they deserve.

Demographics

If I had to describe my political views overall, I would say that I am:
(Response options: 1 = Very liberal to 7 = Very conservative)

With respect to economic matters, I consider myself:

(Response options: 1 = Very liberal to 7 = Very conservative)

With respect to social matters, I consider myself:

(Response options: 1 = Very liberal to 7 = Very conservative)

Gender:

(Response options: Male, Female, Other)

Age:

(Response options: 18 to 97)

Education:

(Response Options: High School or Less, Some College, College Graduate, Some Post-Graduate Education, Masters Degree or Equivalent, Doctoral Degree or Equivalent)

Demographics for Academics Only

Do you work in academia?

(Response Options: No, Yes)

Which of the following best describes your academic career stage:

(Response Options: Research Assistant, Graduate Student, Postdoctoral Scholar, Teaching Faculty, Assistant Professor, Associate Professor, Professor, Professor Emeritus, Other Academic Position)

What year did you receive or do you expect to receive your PhD?

(Open)

If relevant, in what field did you receive your doctoral degree?

(Open)

Do you have tenure at a university?

(Response Options: No, Yes)

How many research articles have you published on the following topics?

	Number of articles
Prejudice and discrimination	
Gender	
Race	
Implicit bias	
Politics	
Evolution	

How many total peer-reviewed academic articles have you published?
(Open)

How many times have you taught a graduate level statistics or methods course?
(Open)

Rate your proficiency in statistics relative to other academics:
(Response Options: 1 = Much lower than average to 9 = Much higher than average)

How familiar are you with research on gender discrimination?
(Response Options: 1 = Not at all familiar to 9 = Extremely familiar)

If the pursuit of truth and the pursuit of social justice came into conflict, which should academia prioritize?
(Response Options: Truth, Social Justice)

Feedback Question (All Participants)

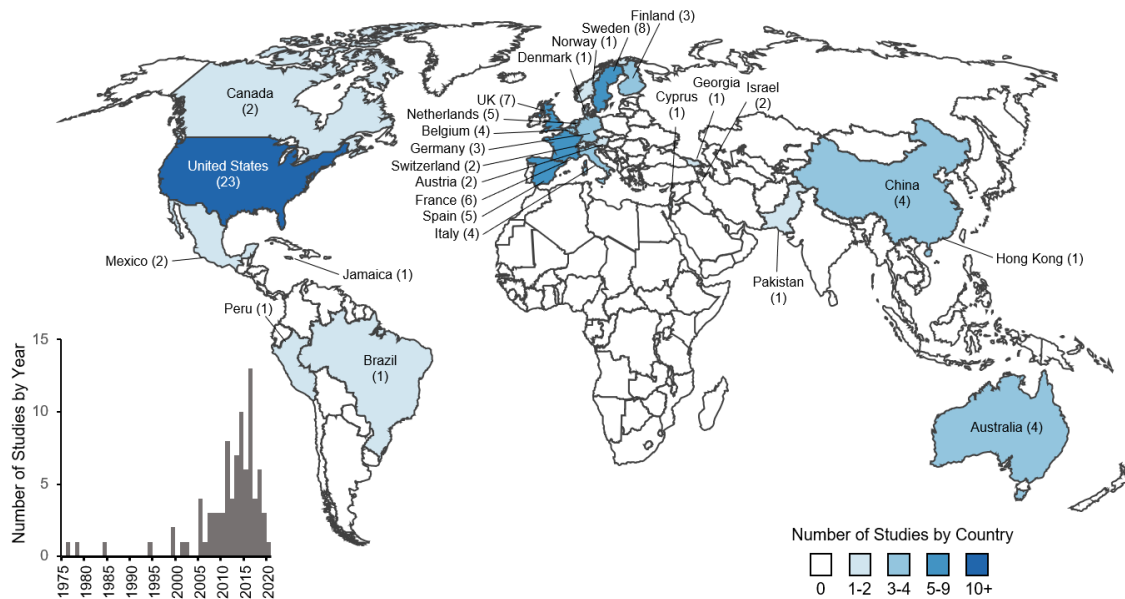
Do you have any feedback for the researchers regarding this survey? For example, did anything seem confusing to you or could the survey be improved in some way? If yes, please write it here.
(Open)

Appendix to Supplement 9: Methods Summary Provided to Forecasters

Gender Meta-Analysis - Methods Summary

We used meta-analytic techniques to examine changes in gender discrimination in hiring decisions over time.

Search. First, we attempted to identify all published and unpublished field audits that examined the effects of candidate gender on hiring decisions. We conducted a systematic search of academic databases, employed backward and forward citation tracing, and identified relevant unpublished work through a search of dissertation databases, public calls for unpublished work, and direct contact with corresponding authors of included studies. Our final sample included 240 effect sizes from 87 studies examining job application outcomes between female and male applicants during the period from 1976 to 2020, representing 373,706 individual job applications across 26 countries and territories (see figure below).



Data extraction. Second, we extracted and processed relevant information from the target articles and reports to create a database for our analyses following our preregistered coding scheme. We extracted objective information at the study level (e.g., gender ratio of authors, publication year, years applications were sent out) and at the effect level (e.g., job type, number of applications sent, callbacks). The stereotypicality of the job(s) applied for were coded by two raters for whether the job(s) was (were) stereotypically female, gender balanced (either a gender-neutral job or mixed set of jobs), or stereotypically male. We also accounted for cross-cultural variation in gender inequality using the United Nation’s Gender Inequality Index (GII) to control for potential geographic shifts in the locations of audit studies over time. 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).

Analysis. Third, we conducted the preregistered and additional exploratory analyses to examine gender discrimination in hiring overall, by job type, and across time. Our outcome variable to assess discrimination was the odds ratio of female and male job applicants that received callbacks (i.e., request for additional information, interview invites, and/or job offers). This was calculated such that a decrease in the odds ratio over time represents a decrease in discrimination against female job candidates. We then conducted a series of multilevel meta-regressions regressing the effect size on key study characteristics (study year, Gender Inequality Index, study design complexity, and author gender composition). Meta-regression is a procedure comparable to standard regression, except that our variables were measured at the study (or effect) level rather than at the individual applicant level.

S12. Study 2: Detailed Report of the Forecasting Results

In the forecasting component of the project, we examined the extent to which scientists and laypeople can predict the results of the meta-analysis of field audits of gender discrimination in selection decisions. We collected forecasts from two groups: scientists primarily from the social and behavioural sciences, and a nationally representative layperson sample from the United States recruited through Prolific. Participants for the academic sample were recruited through social media, professional listservs, and direct email. In line with our preregistration, the sample also contains data from a pilot data collection conducted with graduate students from INSEAD. Both surveys were completed in November/December 2021.

Methods

Participants

The layperson sample 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.

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 are 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); 2 participants did not respond.

Forecasting task

The forecasters separately predicted the meta-analytic results for female-typed jobs, and for male/neutral-typed jobs combined. For each type of job, forecasters predicted the average results of audit studies from the following four time periods: 1976-1986, 1987-1997, 1998-2008 and 2009-2020 and the overall results. For each forecasting question they provided the callback rates (in percentages) separately for female and male candidates respectively, which were then

converted to log odds ratios and compared to the observed log odds ratios from the meta-analysis. See Supplement 10 for the complete survey materials.

Results

Forecasted results are shown in Figure 7 in the main text, alongside the realized effect sizes from the meta-analysis of hiring audits. 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/neutral-typed jobs, and higher callback rates for female candidates (relative to male candidates) for female-typed jobs. The strength of this effect was expected to decline from the earliest to the most recent time period. Laypeople expected stronger effects compared to academics.

For all forecasted log odds ratios, the mean is statistically significantly different from zero (one-sample t-tests), and is statistically significantly different from the observed effects (two-sample z-tests). The p-values are shown in Table S6-1 and Table S6-2.

Table S6-1. Observed and forecasted log odds from the academic sample, together with the p-value from testing the forecasts against a mean of zero, and against the observed effect.

Label	Job type	Time	Observed		Forecasts from academics				
			log odds	se	mean	se	p-value t-test	z-value	p-value z-test
M-T1	Male/Neutral	1976-1986	0.27	0.15	2.10	0.09	< 2.2E-16	-10.40	< 2.2E-16
M-T2	Male/Neutral	1987-1997	0.51	0.26	1.56	0.07	< 2.2E-16	-3.90	9.75E-05
M-T3	Male/Neutral	1998-2008	-0.08	0.05	1.11	0.06	< 2.2E-16	-16.10	< 2.2E-16
M-T4	Male/Neutral	2009-2020	-0.04	0.03	0.72	0.05	< 2.2E-16	-13.66	< 2.2E-16
M-ALL	Male/Neutral	Pooled	-0.03	0.03	1.15	0.06	< 2.2E-16	-18.98	< 2.2E-16
F-T1	Female typed	1976-1986	-0.02	0.22	-1.63	0.11	< 2.2E-16	6.63	3.36E-11
F-T2	Female typed	1987-1997	NA	NA	-1.30	0.08	< 2.2E-16	NA	NA
F-T3	Female typed	1998-2008	-0.42	0.11	-0.99	0.06	< 2.2E-16	4.38	1.17E-05
F-T4	Female typed	2009-2020	-0.23	0.06	-0.69	0.06	< 2.2E-16	5.58	2.42E-08
F-ALL	Female typed	Pooled	-0.28	0.05	-1.01	0.07	< 2.2E-16	8.68	< 2.2E-16

Table S6-2. Observed and forecasted log odds from the U.S. nationally representative sample, together with the p-value from testing the forecasts against a mean of zero, and against the observed effect.

Label	Job type	Time	Observed		Forecasts from US laypersons				
			log odds	se	mean	se	p-value t-test	z-value	p-value z-test
M-T1	Male/Neutral	1976-1986	0.27	0.15	3.52	0.11	< 2.2E-16	-17.38	< 2.2E-16
M-T2	Male/Neutral	1987-1997	0.51	0.26	2.56	0.09	< 2.2E-16	-7.47	8.19E-14
M-T3	Male/Neutral	1998-2008	-0.08	0.05	1.76	0.07	< 2.2E-16	-21.75	< 2.2E-16
M-T4	Male/Neutral	2009-2020	-0.04	0.03	1.11	0.06	< 2.2E-16	-17.96	< 2.2E-16
M-ALL	Male/Neutral	Pooled	-0.03	0.03	1.79	0.07	< 2.2E-16	-24.39	< 2.2E-16
F-T1	Female typed	1976-1986	-0.02	0.22	-2.43	0.14	< 2.2E-16	9.38	< 2.2E-16
F-T2	Female typed	1987-1997	NA	NA	-1.87	0.10	< 2.2E-16	NA	NA
F-T3	Female typed	1998-2008	-0.42	0.11	-1.40	0.07	< 2.2E-16	7.22	5.07E-13
F-T4	Female typed	2009-2020	-0.23	0.06	-1.10	0.06	< 2.2E-16	10.23	< 2.2E-16
F-ALL	Female typed	Pooled	-0.28	0.05	-1.52	0.08	< 2.2E-16	13.66	< 2.2E-16

All analyses for which we had sufficient data were conducted as outlined in the pre-analysis plan (see Supplement 9). However, in an oversight we did not include a forecasting question asking for callback rates across all jobs regardless of stereotypicality (male-typed, neutral-typed, female-typed), and we also neglected to include a binary question on academics identifying as gender researchers. As a result, Primary Hypotheses 1a and 1b (involving overall gender discrimination regardless of job type) could not be tested due to a lack of forecasting data. We were able to examine whether gender researchers are more accurate forecasters using an alternative item asking respondents whether they had ever published a peer-reviewed article on gender (see below). In addition, we have no observed result for female-typed jobs for the second time period since there were no field audits of gender discrimination for such positions during that specific span of years. The affected hypothesis tests are adjusted for cases of unavailable data. Finally, we obtained responses of 0 and 100 for some of the callback rates, which gives infinite log odds ratios. Such responses were capped at 1% and 99% to be able to analyze the data without excluding the most extreme datapoints.

As pre-registered we use both the more conservative significance threshold of $p < .005$ proposed by Benjamin et al. (2018) and the traditional threshold for statistical significance of $p < .05$ in interpreting our results (and all the tests are based on two-sided p-values). Also as pre-registered, only forecasters that answered all forecasting questions are included in the analyses.

Note that below we refer to numbered hypotheses for the forecasting survey as outlined in the preregistered analysis plan for the forecasting analyses (see Supplement 9). These are to be distinguished from the numbered hypotheses associated with the meta-analysis of field audits (Study 1 in the main manuscript).

Primary Forecasting Hypothesis 2a: *Forecasters believe that discrimination against women relative to men is greater in male-typed plus neutral-typed jobs than in female-typed jobs.*

This hypothesis is tested using a paired t-test to compare the forecasters' log odds ratios for male/neutral typed 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, $t(311) = 21.2$, $p < 2.2e-16$, $d = 1.97$) and the layperson sample (mean of differences: 3.31, $t(498) = 29.1$, $p < 2.2e-16$, $d = 2.04$). Note that this is because the forecasted mean log odds ratio for male/neutral typed jobs is larger than zero (higher callback rates for males, see Table S6-1 and S6-2 for the two samples) and the forecasted mean log odds ratio for female-typed jobs is smaller than zero (higher callback rates for female candidates).

Primary Forecasting Hypothesis 2b: *Forecasters under/overestimate the meta-analytic difference in gender discrimination between male-typed plus neutral-typed jobs and female-typed jobs.*

To test this hypothesis, we estimate for each forecaster the difference between the log odds ratios for male/neutral typed jobs for the entire time period and the log odds ratios for female-typed jobs for the entire time period. The mean difference and standard error of this difference is the

used in a z-test to compare against the difference between the difference in the meta-analytic estimates for male/neutral typed jobs and female-typed jobs over the entire time period. We find that differences in discrimination are overestimated in both the academic sample (mean of difference: 1.91, $z = 16.3$, $p < 2.2\text{e-}16$) and the layperson sample (mean of difference: 3.07, $z = 24.0$, $p < 2.2\text{e-}16$). In both cases, the forecasted differences are much larger than the observed differences.

Primary Forecasting Hypothesis 3a: *Forecasters believe that discrimination against women relative to men has decreased over time for male-typed plus neutral-typed jobs.*

This hypothesis is tested using a paired t-test to compare the forecasters' log odds ratios for male/neutral typed 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, $t(311) = 19.0$, $p < 2.2\text{e-}16$, $d = 1.04$; mean of differences in the layperson sample: 2.41, $t(498) = 27.1$, $p < 2.2\text{e-}16$, $d = 1.20$).

We also observe that forecasters expect discrimination against men for female-typed jobs to decrease over time (mean of differences in the academic sample: -0.95, $t(311) = 11.5$, $p < 2.2\text{e-}16$, $d = .62$; mean of differences in the layperson sample: -1.34, $t(498) = 11.6$, $p < 2.2\text{e-}16$, $d = -.57$). Note that these two tests are follow-up tests that were not pre-registered.

Primary Forecasting Hypothesis 3b: *Forecasters under/overestimate the meta-analytic reduction in gender discrimination over time for male-typed plus neutral-typed jobs.*

To test this hypothesis, we calculate for each forecaster the difference between the log odds ratios for male/neutral typed jobs for the first time period and the log odds ratios for the last time period. The mean difference and standard error of this difference is then used in a z-test to compare against the difference between meta-analytic estimates for male/neutral typed jobs for the first and last time period. In both samples, the meta-analytic reduction is overestimated (academic sample mean of difference: 1.07, $z = 6.4$, $p = 1.9\text{e-}10$; layperson sample mean of difference: 2.11, $z = 12.0$, $p < 2.2\text{e-}16$). Note that although the expected reduction in discrimination over time is much larger than the observed reduction, even for the most recent time period forecasters expect more discrimination than is observed.

Secondary Forecasting Hypothesis 1a: *Forecasters believe that men experience more positive job application outcomes than women for male-typed plus neutral-typed jobs.*

This hypothesis is tested by testing the forecasted log odds ratios for male/neutral typed jobs against zero in a one-sample t-test. In both samples, we find a statistically significant difference such that forecasters believe male candidates receive more callbacks than female candidates for

stereotypically male-typed and neutral-typed jobs (see Tables S11-1 and S11-2, row 5 - label M-ALL - for mean and p-values from the one-sample t-test).

Secondary Forecasting Hypothesis 1b: *Forecasters under/over estimate the meta-analytic gender discrimination for male-typed plus neutral-typed jobs.*

To test this hypothesis, a z-test is used to compare the mean of the forecasted log odd ratios for male/neutral typed jobs against the estimated discrimination from the meta-analysis. We find that forecasters overestimate the meta-analytic effect sizes for discrimination in both samples. For forecasted and observed effects and p-values from the two-sample z-tests see Table S6-1 and S6-2 row 5 – label M-ALL).

Secondary Forecasting Hypothesis 2a: *Forecasters believe that women experience more positive job application outcomes than men for female-typed jobs.*

This hypothesis is tested by testing the forecasted log odds ratios for female-typed jobs against zero in a one-sample t-test. In both samples, we find a statistically significant difference such that forecasters believe female candidates receive more callbacks than male candidates for stereotypically female-typed jobs. Mean and p-values from the one-sample t-test are given in Tables S6-1 and S6-2, row 10 - label F-ALL.

Secondary Forecasting Hypothesis 2b: *Forecasters under/over-estimate the meta-analytic gender discrimination for female-typed jobs.*

To test this hypothesis, a z-test is used to compare the mean of the forecasted log odd ratios for female-typed jobs against the estimated discrimination from the meta-analysis. We find that forecasters overestimate the meta-analytic effect sizes in both samples. In other words, forecasters anticipated relatively greater discrimination against male candidates for female-typed jobs than was actually observed (see Tables S6-1 and S6-2, row 10 - label F-ALL - for forecasted and observed mean and p-values from the two-sample z-test).

Secondary Forecasting Hypothesis 3: *There is a positive association between the predictions (beliefs) 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/neutral typed 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. We therefore have seven observations per forecaster. We

include individual fixed effects in the OLS regression and we cluster 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 outcomes for both the sample of academics (coefficient = .09, $t = 17.2$, $p < 2.2\text{e-}16$) and the layperson sample (coefficient = .06, $t = 34.6$, $p < 2.2\text{e-}16$). Note that while forecasters expect much larger effects than actually emerged in the meta-analysis, there is a positive correlational relationship between forecasts and observed outcomes. The coefficient of about .09 in the academic sample and 0.06 in the U.S. layperson sample mean that an expected difference in log odds ratios of 1 translates onto an observed change of less than .1.

Secondary Forecasting Hypothesis 4: The accuracy of predictions differs between scientists and 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 first estimate the mean squared prediction error of each forecaster for the 9 verifiable predictions and then test if this differs between scientists and laypeople using an independent samples t-test. We find that the mean error is significantly smaller for the academic forecasters compared to the laypeople sample (means 2.86 vs. 7.27, $t(803) = -10.3$, $p < 2.2\text{e-}16$). This is because laypeople gave more extreme and therefore less accurate estimates than the academics (see Figure S11-1).

Secondary Forecasting Hypothesis 5: The accuracy of predictions differs for female-typed jobs and male-typed plus neutral-typed jobs.

For this test we construct two individual level variables. The first of these variables is the mean squared prediction error of each forecaster for their predictions for the female-typed jobs, and the second is the mean squared prediction error of each forecaster for the five predictions of the male-typed plus neutral-typed jobs. We then carry out a paired t-test of these two variables to test if the mean squared prediction error differs for the predictions of female-typed jobs and the predictions of neutral-typed and male-typed jobs.

We do not find a statistically significant difference in accuracies in the academic sample (mean squared error of 2.96 for predictions for male/neutral typed jobs vs. 2.74 for female-typed jobs, $t(311) = 0.87$, $p = 0.38$). In the lay sample, forecasts for female-typed jobs were significantly more accurate than for male/neutral-typed jobs (mean squared error of 7.92 for male-type and gender-balanced job predictions vs. a mean squared error of 6.46 for female-typed jobs, $t(498) = 3.81$, $p = .00016$). More specifically, laypeople were more likely to overestimate discrimination against female candidates for male-typed and neutral-typed jobs than to overestimate biases in selection against male candidates for female-typed jobs.

Note that the analyses below were pre-registered in advance, but are labelled as exploratory due to the large number of variables and tests involved.

Exploratory Forecasting Analysis 1: Does the accuracy of predictions vary with political orientation?

Political orientation of the forecaster is estimated as the average of the forecasters' responses to three 7-point scale survey questions about their overall, social, and economic political orientation. The scale of each attitude question is measured from "very liberal" (-3) to "very conservative" (3), with 0 as the neutral midpoint. The three political orientation questions are phrased as: "If I had to describe my political views overall, I would say that I am..."; With respect to economic matters, I consider myself..."; With respect to social matters, I consider myself...". We estimate an OLS regression with the mean squared prediction error of each forecaster as the dependent variable (see secondary forecasting hypothesis 4 for definition) and the political orientation variable as the independent variable. The OLS regression is estimated with robust standard errors, and the test is carried out as a t-test on the regression 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 = .048, $t = .25$, $p = .80$) or in the layperson sample (coefficient = -.33, $t = 1.52$, $p = .13$).

Exploratory Forecasting Analysis 2: Does the accuracy of predictions vary with system justification scores?

Individual differences in system justification are estimated as the average response of a forecaster to eight 7-point scale survey questions (e.g., "In general, society is fair"). We estimate an OLS regression with the mean squared prediction error of each forecaster as the dependent variable and the system justification scale score as the independent variable. The OLS regression is estimated with robust standard errors, and the test is carried out as a t-test on the regression coefficient of the system justification variable in the OLS regression. For the academic sample, we find that individual differences in system justification are associated with reduced error in predictions (coefficient = -.46, $t = -2.48$, $p = .014$). For the U.S. layperson sample, with increasing endorsement of the items on the system justification scale, the error likewise decreases (coefficient = -0.83, $t = -2.65$, $p = .008$). 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). There is thus suggestive evidence that academics and laypersons who were less generally egalitarian (i.e., higher in system justification) were more accurate in their forecasts about gender discrimination (i.e., made fewer errors). Although this analysis was pre-registered, we consider these results exploratory given the large numbers of secondary analyses carried out. Further, only the lax, not the strict criteria for statistical

significance that we set out beforehand was met. Thus, the finding that less egalitarian forecasters were more accurate needs confirmation in further studies to carry more weight.

Exploratory Forecasting Analysis 3: Does the accuracy of predictions vary with gender system justification?

Gender system justification is estimated as the average response of a forecaster to eight 7-point scale survey questions (e.g., “Everyone (male or female) has a fair shot at wealth and happiness”). We estimate an OLS regression with the mean squared prediction error of each forecaster as the dependent variable and the gender system justification variable as the independent variable. The OLS regression is estimated with robust standard errors, and the test is carried out as a t-test on the regression coefficient of the system justification variable in the OLS regression. 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 $p < .05$ cutoff is used, but not when the more conservative $p < .005$ cutoff 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 cutoff is used. There is thus evidence that forecasters who were less gender egalitarian made more accurate forecasts about gender discrimination. This evidence is suggestive for the academic sample, and comparatively much stronger for the lay sample. Although pre-registered, we consider this an exploratory test given the large number of secondary analyses reported here.

Exploratory Forecasting Analysis 4: Does the accuracy of predictions differ between scientists with backgrounds studying gender bias and other scientists?

In an oversight, we did not include a binary (yes/no) self-identification question asking academic forecasters whether they considered themselves a gender researcher. However, we did include a series of items asking “How many research articles have you published on the following topics?” with “Gender” as one of the topics. We therefore tested this research question by categorizing forecasters who had published at least 1 paper on gender as a gender researcher ($N = 132$) and categorizing forecasters without published papers on gender as non-gender researchers ($N = 168$). 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, $t(250) = .037$, $p = .71$). Forecasters who did not provide a response ($N = 11$) were excluded from this analysis.

Exploratory Forecasting Analysis 5: We will carry out the tests in primary hypothesis 2a and 2b separately for each of the four time periods.

In both the academic and lay samples, we find that forecasters believe for each time period that discrimination of women relative to men is larger in male-typed plus neutral-typed jobs than in female-typed jobs. The effects are statistically significant for each time period and in both samples ($p < .001$ for all tests; t -values omitted). We also find that differences in discrimination are overestimated in both the academic sample and the U.S. lay sample, for all available time periods ($p < .001$ for all tests; t -values omitted).

Exploratory Forecasting Analysis 6: *We will carry out the tests in secondary hypothesis 1a and 1b separately for each of the four time periods.*

Forecasters in both samples believe that in each of the time periods, men experience more positive job application outcomes than women for male-typed and neutral-typed jobs. The forecasters overestimate the observed discrimination for each time period. Differences are all statistically significantly different, and p -values are given in Tables S11-1 and S11-2.

Exploratory Forecasting Analysis 7: *We will carry out the tests in secondary hypothesis 2a and 2b separately for each of the four time periods.*

Forecasters in both samples believe that in each of the time periods, female candidates experience more positive job application outcomes than male candidates with regards to stereotypically female-typed jobs. The forecasters overestimate observed discrimination for each time period in a sense that the forecasted log odds ratio is more negative than the observed meta-analytic estimate. Differences are all statistically significantly different, and p -values are given in Tables S11-1 and S11-2.

Exploratory Forecasting Analysis 8: *We will compare the mean forecast between the sample of scientists and the sample of laypeople for each of the forecasting questions.*

Using independent sample t -tests to compare the forecasts from the academics and layperson, we find for each of the forecasting questions statistically significant differences between the means. For the forecasts for male/neutral typed jobs, we find that laypeople provided more positive estimates than scientists, and for the female-typed jobs more negative estimates than scientists. More positive estimates mean a greater expectation of bias in favor of male candidates, whereas negative estimates reflect an expected bias in favor of female candidates. All differences are statistically significant (with $p = .0001$ or smaller; t -values omitted). Thus, laypeople consistently overestimate gender discrimination more than academics.

Potential caveat and additional robustness test

We observed that in a number of cases, the two paired responses (female candidate and male candidate callback rates) sum up to 100%. This suggests that some participants might have interpreted the question incorrectly and provided forecasts for the composition of candidates

receiving callbacks, rather than callback rates. This composition depends on callback rates, but also the initial composition of applicants. This affects about 20% of the academic sample and 60% of the U.S. representative sample of laypersons. The forecasts differ systematically between forecasters whose rates add up to 100 for each pair of callback rates, and forecasters for whom this is not the case. In both samples, the log odds ratios are more extreme when the callback rates add up to 100.

Therefore, as an additional robustness test, we repeated the above analyses after excluding all forecasters whose rates add up to 100 for each pair of callback rates. We find that all statistically significant effects remain statistically significant, with the following exceptions:

- In **Exploratory Forecasting Analysis 2** (Does the accuracy of predictions vary with system justification scores?), the effect in the academic sample is no longer statistically significant (coefficient = $-.27$, $t = -1.62$, $p = .11$, compared to the original analysis with coefficient = $-.46$, $t = -2.48$, $p = .01$).
- In **Exploratory Forecasting Analysis 3** (Does the accuracy of predictions vary with gender system justification scores?), the effect in the academic sample is no longer statistically significant (coefficient = $-.16$, $t = -0.95$, $p = .34$, compared to the original analysis with coefficient = $-.45$, $t = -1.98$, $p = .049$).
- In **Exploratory Forecasting Analysis 8** (comparison of the mean forecast between the sample of scientists and the sample of laypeople for each of the 11 forecasting questions), the differences for female-typed jobs in time periods 1-3 no longer meet the criterion for statistical significance of $p < .005$, but still meet the less stringent p -value criterion of $p < .05$.

Reference S11

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, 6–10. doi:10.1038/s41562-017-0189-z

S13. Supplementary References

- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E., Berk, R., Bollen, K. A., Brembs, B., Brown, L., & Camerer, C. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6-10.
- Coburn, K. M. & Vevea, J. L. (2016). Package ‘weightr’. *Estimating Weight-Function Models for Publication Bias*.
- 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.
- Hartung, J. & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, 20(12), 1771-1782.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600-2606.
- Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, 23(10), 815-818.
- 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.
- 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.
- Raudenbush, S. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis*: 295-315. New York, NY: Russell Sage Foundation.
- Sidik, K. & Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine*, 21(21), 3153-3159.
- Stanley, T. D. & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1), 60-78.
- van Aert, R. (2021). puniform: Meta-analysis methods correcting for publication bias (Version 0.2.4).
- Vevea, J. L. & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3), 419-435.
- Viechtbauer, W. 2007. Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26(1), 37-52.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48.
- Viechtbauer, W. & Cheung, M. W. L. 2010. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, 1(2), 112-125.
- 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.
- Wang, M. & Bushman, B. (1998). Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods*, 3(1): 46-54.