Sensitivity and Response Bias in Detecting Gender Discrimination

Marie Jakob[1], Anat Shechter[1], Jimmy Calanchini[2], and & Karl Christoph Klauer[1]

[1] University of Freiburg

[2] University of California Riverside

Preprint submitted for publication on July 4th 2024.

Author Note

This project has received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 503990132. Marie Jakob additionally received support from the research training group Statistical Modeling in Psychology (SMiP; grant GRK 2277), funded by the German Research Foundation (DFG).

This paper was written as a reproducible document in RMarkdown using the papaja package (Aust & Barth, 2023) and includes the code for analyses and figures. The RMarkdown document, all additional analysis scripts, and all data collected for this project are openly available at https://osf.io/6cgxv/.

The authors made the following contributions. Marie Jakob: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing; Anat Shechter: Conceptualization, Methodology, Writing - Review & Editing; Jimmy Calanchini: Conceptualization, Methodology, Writing - Review & Editing; Karl Christoph Klauer: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Software, Supervision, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Marie Jakob, Engelbergerstraße 41, 79085 Freiburg im Breisgau, Germany. E-mail: marie.jakob@psychologie.uni-freiburg.de

Abstract

Judgments about the presence or absence of discrimination are generally made in a context of considerable uncertainty. Here, we apply Signal Detection Theory, a framework that models decisions under uncertainty, to attributions to discrimination in order to differentiate between sensitivity (i.e., the ability to differentiate between situations with and without cues to discrimination) and response bias (i.e., a relative tendency to make liberal rather than conservative judgments about the presence of discrimination) in such judgments. To do so, we developed a new paradigm and used it to examine participants' attributions to gender discrimination by manipulating the gender of victims and observers and the harm of the action (Experiment 1), the direction of gender discrimination (Experiment 2), the discrimination context (Experiment 3) and participants' decision task (Experiment 4). Participants' response bias was consistent with cultural stereotypes about gender discrimination and was affected by the actual base rate of discrimination in the experimental context which they picked up explicitly only in the context of discrimination. The harm implied by the action in question affected participants' response bias such that they were more liberal in cases in which victims were harmed, with small corresponding differences in sensitivity. Our flexible and broadly applicable new paradigm opens up new possibilities for further research on attributions to discrimination and their analysis in terms of sensitivity and response bias.

Public Significance Statement

We investigated how people detect gender discrimination and introduce a new method to separate a person's ability to detect discrimination from their tendency to assume that discrimination is present. Employing this method in four studies, our results show that people have a stronger tendency to judge situations as being discriminatory when they match with cultural stereotypes about gender discrimination. Their ability to accurately detect discrimination, however, was unrelated to cultural stereotypes. Similarly, people's

tendency to judge situations as discriminatory increased with the frequency and severity of discrimination. Our new method opens up new possibilities for future research on discrimination by providing a detailed perspective on people's judgments in terms of their ability and decision tendencies.

Keywords: Discrimination; Gender Bias; Signal Detection Theory; Sexism; Prejudice

Word count: 14858

Sensitivity and Response Bias in Detecting Gender Discrimination

Introduction

Gender discrimination, that is, the unfair or prejudicial treatment of people and groups based on their gender (American Psychological Association, 2019), is a longstanding and notorious issue with detrimental consequences both for the people affected by it (Crocker & Major, 1989; Jones et al., 2016; Schmitt et al., 2014) and society as a whole (Blau & Kahn, 2016). In order to combat gender discrimination, being able to *detect* it is a necessary prerequisite; however, the evidence for or against discrimination in specific situations is often ambiguous (e.g., Barreto & Ellemers, 2015; Crocker & Major, 1989; F. Crosby et al., 1986; Salvatore & Shelton, 2007). According to the definition above, attributions to discrimination require two components: (1) The treatment of an individual is judged to be unjust, and (2) the cause for that treatment is attributed to the individual's group membership. Ambiguity regarding (1) is due to the fact that standards for what constitutes fair treatment are often not available. If, for instance, we observe a car salesperson extensively explaining color schemes of different car brands to a female customer while leaving out technical aspects, we presumably do not have a clear-cut standard on how the salesperson is supposed to act. If such standards are not available, discrimination can still be assessed through unjust *differential* treatment of an individual (Major, Quinton, et al., 2002). However, the contrast cases required to diagnose differential treatment are also often not available. In the just-described example, for instance, we are unlikely to observe the same salesperson in a comparable situation involving a male customer. Moreover, the attribution of the behavior to group membership is complicated by the fact that group membership and other variables are often confounded in real life. For example, a woman might not get promoted to a leadership position because she is a woman or because she is less vocal in company meetings. Finally, strong normative pressure exists in society against prejudice and discrimination; as a consequence, discrimination is often only expressed in

subtle ways (e.g., Barreto & Ellemers, 2015).

The detection of gender discrimination in particular is even more ambiguous due to the fact that gender discrimination – in contrast to other forms of discrimination – manifests in two forms (Glick & Fiske, 1996): *hostile* sexism, characterized by overtly negative behaviors (e.g., cat-calling, sexist language etc.), and in *benevolent* sexism, characterized by attitudes and behaviors towards women that seem positive in tone but nevertheless imply inferiority to men (e.g., asking only female co-workers to organize an office party because women are „naturally better" at handling social events). Both forms of discrimination contribute to the maintenance of patriarchal structures but benevolent discrimination, due to its superficially positive appearance, often goes unrecognized (Barreto & Ellemers, 2005). Given this backdrop, the attribution of an outcome to discrimination in general and to gender discrimination in particular regularly occurs in a context of considerable uncertainty.

Sensitivity and Response Bias

Judgments under uncertainty can be characterized by two distinct components: Sensitivity and response bias. In the context of attributions to discrimination, sensitivity describes the ability to accurately differentiate between situations with and without cues to discrimination. Response bias, on the other hand, describes the tendency to make an attribution to discrimination rather than other causes; in other words, it refers to the extent to which judgments about the presence or absence of discrimination are made according to a liberal or conservative decision threshold. Disentangling differences in sensitivity and response bias is the goal of signal detection theory (SDT; Kellen & Klauer, 2018; Green & Swets, 1966). The central assumption of SDT is that decisions under uncertainty are based on noisy subjective evidence (in this case, subjective evidence for discrimination) and a decision threshold. If (and only if) in a given situation the evidence for discrimination surpasses this threshold, the situation is judged to constitute an instance of discrimination. The decision threshold is a measure of response bias, whereas the difference between mean

evidence in situations with versus without cues to discrimination (i.e., *signal* and *noise*, respectively) is a measure of sensitivity.

By separating these two judgment components, the application of SDT allows one to characterize empirical findings as sensitivity effects, response bias effects, or combinations thereof. Such a characterization may in turn inform theories on attributions to discrimination by clarifying mediational pathways. To illustrate, consider the finding that women are more likely to attribute observed negative outcomes concerning themselves or other women to discrimination than men (Major et al., 2003; Rodin et al., 1990; Stangor et al., 2002; see also J. R. Crosby, 2015). This effect could reflect differences in sensitivity and/or differences in response bias. Two factors may enhance the sensitivity of women: In line with an idea going back to Allport's (1954) vigilance perspective, women, as members of a marginalized group, might invest more attentional resources to monitor events for signs of discrimination (Kaiser et al., 2006). In addition, more frequent exposure to discrimination may have made women experts in detecting cues to gender discrimination. On the other hand, gender differences in response bias are also plausible: Frequent exposure to acts of discrimination may raise women's perceived base rate of discrimination, and base rates are a factor known to affect response biases in many domains (Kellen & Klauer, 2018). People from marginalized groups may also sometimes be motivated to make attributions to discrimination as a means to deflect blame for negative outcomes away from them or their ingroup (Adams et al., 2006; Major, Quinton, et al., 2002), leading to a more liberal response bias. Finally, heightened vigilance might make discrimination a highly accessible construct for women (Higgins, 1996), raising not only the likelihood that they interpret cues to discrimination as such, but also the likelihood that they construe innocuous or ambivalent situational features as cues to discrimination. Higher perceived base rates, motivated attributions, and heightened vigilance might therefore lead to a more liberal response bias for women, whereas stronger monitoring and increased expertise through exposure might increase their sensitivity. An empirical SDT investigation would help clarify whether women's increased likelihood of attributing negative

outcomes to discrimination is a sensitivity effect mediated by their ability to perceive discrimination, a response-bias effect mediated by their propensity to make attributions to discrimination, or a combination of both.

In an influential paper, Barrett and Swim (1998) used SDT as a theoretical framework within which to conceptualize attributions to discrimination. They suggested that people adapt their response bias based on subjective costs of different types of judgment errors and hypothesized about different factors affecting people's sensitivity and response bias. For instance, they speculated that a person's belief about base rates of discrimination against different social groups influences their response bias whereas their general knowledge about social interactions or prejudice affects their sensitivity. Meanwhile, although the work of Barrett and Swim (1998) was highly influential, researchers seldomly moved from considering SDT as a *theoretical* framework within which to construe and hypothesize about mechanisms behind empirical findings to actually testing such hypotheses using *empirical* applications of SDT in this context.[1]

The Bias Detection Task

Most research paradigms that are used to study attributions to discrimination fall into one of two categories. One line of research relies on vignettes that describe situations potentially involving discrimination and asks participants to assess the extent to which the described events reflect discrimination (Inman & Baron, 1996; König & Heine, 2023; O'Brien et al., 2008; Phillips & Jun, 2022; Simon et al., 2019; e.g., in bogus selection procedures or tests, Swim et al., 2003). In another line of research, participants are asked to judge actual events involving discrimination against themselves or another person (Barreto & Ellemers, 2015; Inman, 2001; O'Brien et al., 2008; Stroebe et al., 2009). In both approaches,

[1] The only exception to this that we are aware of is the work of König and Heine (2023) who used SDT in conjunction with Receiver Operating Characteristic analyses in this context to measure people's accuracy independently of their response biases. They used a small number of vignettes that were specifically tailored to their design and analyzed aggregated data. In contrast, our goal was to develop a very general paradigm that can be adapted to different types of discrimination and contexts and that allows one to estimate individual SDT parameters.

participants are usually presented with few trials (in vignette studies) or even only a single trial (in studies involving mock discrimination) in which cues to discrimination are present (i.e., signal trials); corresponding trials without cues to discrimination (i.e., noise trials) are typically not shown (with the exception of Swim et al. 2003, Inman, 1996 and König & Heine, 2023). These features make both of these paradigms ineligible for the application of SDT because the estimation of an SDT model requires a large number of both signal and noise trials (e.g., Kellen & Klauer, 2018).

Here, we introduce a new paradigm – the Bias Detection Task (BDT)[2] – which satisfies these requirements and thereby allows one to estimate participants' sensitivity and response bias based on their responses. Throughout this paper, the BDT is set in the context of pay raise decisions in a fictional marketing company.[3] Employees in this company are evaluated for possible pay raises by a committee of supervisors based on scores indicating their performance regarding three job-relevant criteria. Typically, the panel of supervisors bases their decision on the scores, but additional information is said to be available to them (e.g., the CV, employment history, etc. of the employee in question) and may in special cases weigh in on the decision. This context introduces ambiguity about the causes of specific outcomes and establishes discretionary latitude for the committee which may lead to biased decisions. Participants first complete a learning phase in which they familiarize themselves with how scores are typically translated into pay raise decisions. The purpose of the learning phase is to establish a standard for deservingness against which participants can judge the committee decisions as fair or unfair. In the subsequent evaluation phase, participants are informed about complaints within the company that the committee has made a number of biased decisions and that therefore, the company has decided to organize an independent

---

[2] The term "gender bias" is used inconsistently in the literature, sometimes only including biased *actions* based on someone's gender and sometimes also referring to biased *evaluations* or attitudes and stereotypes. In the context of this task, we use the terms "bias" and "discrimination" interchangeably because participants have to make judgments about *actions*, not directly about gender bias on the level of attitudes or stereotypes.

[3] The rationale of the BDT can be easily applied to other contexts as already exemplified in the studies reported below.

review of the committee decisions. Participants are told that they are involved in a first screening of decisions that is based on basic information for each case (i.e., the scores, the employee's name, a portrait picture and the committee's decision). Their task is to flag cases that are likely biased for further in-depth review. Based on their judgments of the cases as biased or unbiased, participants' response bias and sensitivity for different groups of employees and decisions can be estimated.

The procedure of the BDT is similar to Axt and colleagues' (2018) judgment bias task which is used to measure various social biases and also employs an SDT approach. In that task, participants are also presented with cases showing qualification scores and demographic information of applicants and have to decide whether the applicants should be admitted to an academic honor society. The major difference between these two tasks is that in the BDT, participants judge *given* decisions as biased or unbiased, whereas in the judgment bias task, participants make those decisions *themselves* (i.e., there are no committee decisions, let alone biased decisions). Thereby, the judgment bias task aims to measure *participants' own biases* whereas the BDT assesses differences in participants' *attributions to bias* for given decisions. Accordingly, response bias in the judgment bias task measures a bias towards one or the other decision and sensitivity describes the ability to make accurate decisions. In contrast, response bias in the BDT reflects a general inclination to consider decisions biased and sensitivity reflects the ability to detect bias. In Experiment 4 of this manuscript, we empirically explore differences between responses in the two tasks.

The BDT allows researchers to investigate discrimination based on different social group memberships by manipulating demographic characteristics of the employees (e.g., via different names and portraits). By presenting favorable and unfavorable committee decisions (i.e., a pay raise was either granted or denied) it is possible to investigate attributions to bias in the context of disfavoring as well as favoring discrimination. In addition, the direction of bias can be manipulated through the proportion of cases in which the committee favors or disfavors members of different social groups. Thereby, the BDT permits researchers to

examine attributions to discrimination in cases involving both members of marginalized and privileged groups, extending previous research which mostly focused on discrimination against members of marginalized groups (Major, Quinton, et al., 2002; Swim et al., 2003). In the next section, we discuss our expectations for results patterns in the BDT based on the extant literature.

Attributions to Gender Discrimination in the BDT

Base Rates and Cultural Stereotypes

In the context of gender discrimination, several findings from the literature hint at possible response bias and sensitivity effects in the BDT.[4] One important factor affecting people's attributions to gender discrimination is the base rate of discrimination in a specific context such that when discrimination occurs more frequently, participants make more attributions to discrimination (Barrett & Swim, 1998; Inman, 2001; Major, Quinton, et al., 2002). Correspondingly, SDT results from different domains show that people pick up base rates of different types of trials and shift their response bias accordingly (i.e., more liberal for cases that are shown more often; Kellen & Klauer, 2018]. Thus, in the context of the BDT we expect participants' response bias to reflect the experimentally manipulated base rates of favoring and disfavoring discrimination of male and female employees in the observed committee decisions. In addition to the implemented base rates, cultural stereotypes about gender discrimination in the context of corporate pay raise decisions (e.g., related to the gender wage gap) could affect participants' judgments (O'Brien et al., 2008). Such stereotypes could give rise to corresponding expectations about gender bias in the committee decisions (i.e., that the committee disfavors women and favors men), which have been shown to affect people's attributions to discrimination and have been hypothesized to affect

[4] In the following, we will focus on gender discrimination in the context of a binary understanding of gender. However, we do not mean to imply that there actually are only two genders. Our work merely builds on previous research on gender discrimination focused on men and women and may be extended in future work to include people with gender identities beyond these binary categories.

response bias rather than sensitivity (Barrett & Swim, 1998).

Harm

Previous research also suggests that people consider the harm caused to the victim by the potentially discriminatory behavior when the actor's intent is unclear (Simon et al., 2019; Swim et al., 2003; York, 1989) – as is arguably the case in the BDT where the decisions are made by an anonymous committee. Thus, participants' judgments likely differ between cases in which a pay raise was denied (which causes considerable harm to the employee in question) versus cases in which a pay raise was granted (which does not harm but rather benefits the employee). Such a pattern might again reflect differences in response bias or sensitivity: On the one hand, Barrett and Swim (1998) hypothesized that the intensity of a stimulus increases as harm increases, which reduces ambiguity and could lead to higher sensitivity. On the other hand, they expected the subjective costs of misses (i.e., falsely deciding that no bias / signal is present) and false alarms (falsely deciding that bias / a signal is present) to affect response bias given that in other applications of SDT, those actual costs (as induced by pay-off manipulations, for instance) affect only response bias but not sensitivity. In the context of the BDT, given a "pay raise denied" decision, participants would likely regard the cost of a false alarm (i.e., judging an unbiased decision to be biased) to be lower than the cost of a miss (i.e., judging a biased decision to be unbiased) and shift their response bias accordingly (and vice versa for "granted" decisions).

Favoring and Disfavoring Discrimination

In the BDT, cases in which an undeserved pay raise was granted and a deserved pay raise was denied reflect instances of favoring and disfavoring discrimination, respectively. Recent evidence suggests that participants make fewer attributions to discrimination when presented with cases displaying favoring versus disfavoring discrimination, independent of the harm implied by the decision (Phillips & Jun, 2022). This finding could reflect a more conservative response bias for favoring compared to disfavoring discrimination, mediated by

expectations about discrimination that prioritize disfavoring over favoring discrimination. However, differences in sensitivity are plausible as well: Lay people commonly view discrimination exclusively in terms of disfavoring discrimination (Phillips & Jun, 2022). Consequently, they might have more elaborated knowledge of disfavoring discrimination, increasing their sensitivity to detect it relative favoring discrimination.

Gender Differences

Finally, as elaborated above, men and women may differ in their sensitivity or response bias due to their diverging experiences with gender discrimination. More precisely, heightened vigilance for gender bias, motivated attributions, and a higher perceived base rate of gender discrimination might lead to a more liberal response bias for women, whereas stronger monitoring for signs of discrimination and increased expertise regarding gender discrimination might increase their sensitivity. Taken together, these arguments suggest that there might be differences between male and female participants' response bias and sensitivity in the BDT.

The Present Research

To address the questions discussed above, we used the BDT to characterize people's response bias and sensitivity in attributions to gender discrimination. Table 1 provides an overview of the experiments and conditions reported in this manuscript. Data were collected in three waves. In the first wave, we collected data from the BDT in what we call the "Stereotypical Bias Condition" (column "BDT" in Table 1). In the second wave, this condition was replicated and extended by three new conditions labeled "Counter-Stereotypical Bias Condition", "Placement Condition", and "Decision Condition". Participants were randomly assigned to one of these four conditions. In the third wave, data were collected in what we call the "Symmetrical Bias Condition". For reasons of exposition, we discuss the analyses contrasting different subset of these conditions as different experiments, as shown in Table 1.

Experiment 1 introduces our general experimental paradigm and examines male and female participants' response bias and sensitivity in judgments about the presence of (favoring or disfavoring) discrimination of male and female employees. Based on that, Experiment 2 focuses on the role of the direction of gender discrimination by manipulating the committee's overall gender bias (in the Stereotypical Bias, Symmetrical Bias and Counter-Stereotypical Bias Conditions, see Table 1). Experiment 3 explores the impact of the discrimination context on participants' response bias and sensitivity within our paradigm through an additional condition in which the discrimination context is removed (i.e., the Placement Condition, see Table 1). Finally, Experiment 4 implements a deeper analysis of participants' decision processes when making judgments about the presence or absence of bias through the Decision Condition (see Table 1) in which participants have to make pay raise decisions themselves.

Table 1

Overview of all experiments and conditions.

		Random Assignment				
	BDT	Stereotypical Bias Condition	Counter-Stereotypical Bias Condition	Placement Condition	Decision Condition	Symmetrical Bias Condition
Exp. 1	×					
Exp. 2		×	×			×
Exp. 3		×		×		
Exp. 4		×			×	

Note. The Stereotypical Bias Condition constitutes a replication of Experiment 1. We collected data for the four conditions displayed in the middle columns in one joint data collection with random assignment to the conditions to be able to contrast all conditions rigorously with one another.

Experiment 1: Attributions to Gender Discrimination

In Experiment 1, we used the BDT to explore response bias and sensitivity of men and women when judging favoring and disfavoring discrimination of male and female employees in a context exhibiting gender bias in line with cultural stereotypes. To do so, we presented male and female participants with a number of cases concerning male and female employees (i.e., the cases in the BDT showed portraits of men and women together with male and female first names) that were either granted or denied a pay raise. A quarter of all decisions was biased according to gender stereotypes, such that a female employee was denied a pay raise even though she deserved one (i.e., she was disfavored by the committee) or a male employee was granted a pay raise even though he did not deserve one (i.e., he was favored by the committee). Based on the theoretical considerations and empirical findings above as well as a pilot study, we explored the following research questions:

Research Question 1 (RQ 1): Base Rates and Cultural Stereotypes

We expected participants' judgments to depend on the interaction between the employee's gender and the committee decision, such that participants will make more attributions to discrimination when the cases involve a female employee being denied or a male employee being granted a raise than in the remaining cases. Such a pattern would likely reflect cultural stereotypes about gender discrimination in general as well as the base rates implemented in our design. Based on the arguments above, we hypothesized that both factors influence participants' response bias rather than sensitivity, such that they are more liberal in cases in which female employees are denied and male employees are granted a pay raise.

Research Question 2 (RQ 2): Harm

In line with the effects of harm and the differences between people's perceptions of favoring and disfavoring discrimination, we expected differences in participant's judgments of cases in which a pay raise was granted and in which one was denied (i.e., an effect of the committee decision). Based on our pilot results, we hypothesized the committee decision to affect response bias (i.e., a more liberal response bias for cases where a pay raise was denied), but we additionally tested the corresponding effect on sensitivity (i.e., the possibility of higher sensitivity for such cases).

Research Question 3 (RQ 3): Gender Differences

Furthermore, we examined the extent to which male and female participants differ in their judgments: As we discussed above, women may have a more liberal response bias and / or a higher sensitivity compared to men for several reasons. However, our pilot study revealed a surprising trend towards a higher sensitivity among male participants, which we preregistered accordingly.[5] We therefore focus on sensitivity differences in the following experiments but still tested exploratorily for a corresponding response bias effect.

Research Question 4 (RQ 4): Intergroup Effects

Finally, we expected differences in male and female participants' response bias depending on the gender of the employee in question and the committee decision. If people make motivated attributions to discrimination as a means to deflect blame from their ingroup, their response bias in the BDT might reflect such a pattern. Consequently, we expected participants to be relatively more liberal to judge cases as biased in which employees from their gender ingroup were denied or employees of their gender outgroup were granted a pay raise (i.e., a relative ingroup-favoring response bias).

---

[5] We discuss and test potential explanations for that effect throughout the remainder of this manuscript.

Methods

Transparency and Openness

This paper was written as a reproducible document in RMarkdown using the papaja package (Aust & Barth, 2023) and includes the code for analyses and figures. The data were analyzed using R version 4.2.3 (R Core Team, 2023), and the packages lme4 (Bates et al., 2015) version 1.1.32, tidyverse (Wickham et al., 2019) 2.0.0 and emmeans (Lenth, 2023) 1.8.6. All studies reported in this paper were preregistered on the OSF. The preregistrations, the RMarkdown document, additional analysis scripts, all data collected for this project and materials are openly available at https://osf.io/6cgxv/. We obtained an Ethics approval from the ethics committee of the University of Freiburg for all studies within this project before any data were collected.

Sample

250 participants completed the experiment which was conducted as an online experiment on Prolific, as were the other studies reported in this manuscript. The participants of all of our studies had to be White[6] cisgender native-English speakers, be located in the UK, the USA, Ireland, Australia or Canada, be between 18 and 49 years old, have a high school degree, a perfect score for their previous submissions on Prolific (i.e., no researcher on Prolific has rejected one of their participations based on low-quality data) and must have had already participated in at least five studies on Prolific. Three participants were excluded ($n = 2$ because their mean sensitivity was close to zero, indicating that they did not work properly on their task and $n = 1$ because they reported being a different ethnicity than White), resulting in a final sample size of $N = 247$ (122 women). The mean age of the sample was $M = 34.03$ ($SD = 7.51$), ranging from 19 to 49. Our target sample size for all studies of 120 participants per group (with the exception of one condition in Experiment 2) was determined by a power analysis based on effect size estimates from a pilot

[6] The factor race will be considered in future studies.

study (see the preregistration for details).

Design

The experiment implemented a 2 (*case type*: unbiased vs. biased) × 2 (*employee gender*: female vs. male) × 2 (*committee decision*: pay raise granted vs. denied) × 2 (*participant gender*: male vs. female) design; all factors except for *participant gender* were manipulated within participants. The design was unbalanced and not fully crossed: In total, the committee decisions were unbiased in 192 (75 %) of the 256 total cases in which either a deserved pay raise was granted (96 cases) or an undeserved pay raise was denied (96 cases). The remaining 64 (25 %) cases involved biased decisions in which either a deserved pay raise was denied to a female employee (32 cases) or an undeserved pay raise was granted to a male employee (32 cases). Female employees were never granted an undeserved pay raise and male employees were never denied a deserved pay raise. The cases were presented in an individually randomized order.

Procedure

The experiment comprised a practice phase, an evaluation phase and a rating phase. After giving informed consent, participants received instructions about the fictional marketing company and its procedure regarding the pay-raise decisions. They were told that the pay-raise decisions follow a two-step procedure. In the first step, every employee would receive scores on three criteria. Participants received detailed information on the criteria and the meaning of different scores and were told to familiarize themselves with them. Next, participants responded to the single-choice questions of a first comprehension check. For all comprehension checks, participants had two chances to respond correctly to the questions. If they did not do so on their second try, the experiment immediately ended for them. After passing the first comprehension check, participants were informed that in the second step of the pay raise decisions, a committee of supervisors would decide for each employee separately, whether a pay raise should be granted or not. They were told that this decision

was usually made based on the three scores, but that the committee had additional information about the employee's CV and previous employment history at their disposal and that that information could in special cases influence their decision. This procedure was intended to establish a standard for fair pay raise decisions while giving the committee discretionary latitude whereby biases could influence the decisions. Participants then had to answer a second comprehension check.

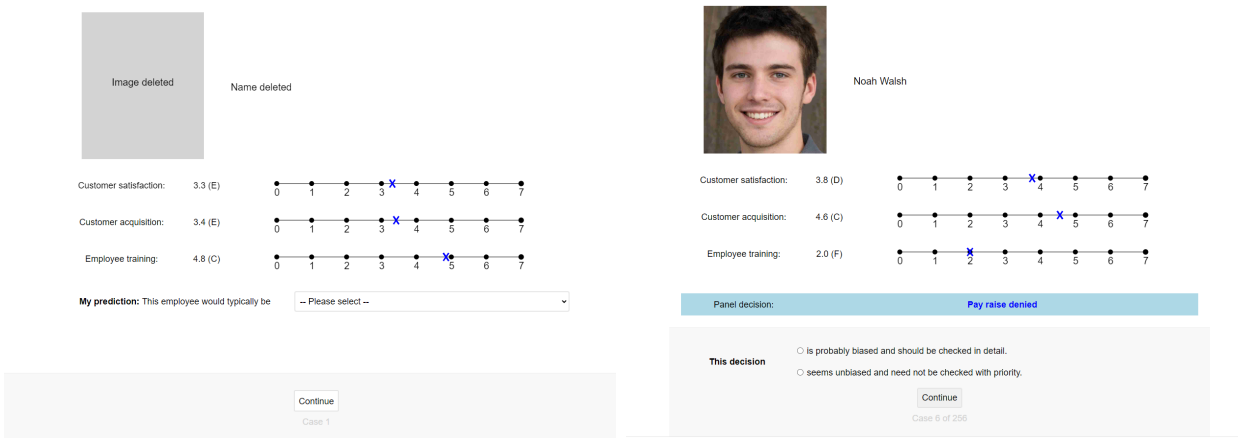

Figure 1

Example trials of the practice (left) and evaluation phase (right).

To familiarize themselves with how the pay raise decisions are typically made, participants completed a practice phase in which they were presented with a number of anonymized cases and had to predict for each case what decision the committee would typically make (see Figure 1, left panel). After they had made their decision, they were shown the actual committee decision for that case. Once participants had predicted 18 out of 20 consecutive decisions correctly, they proceeded to the evaluation phase. Participants were instructed that the company had received complaints about biased decisions, such that some employees who deserved a pay raise were denied and other employees who did not deserve a pay raise were granted one. They were further told that the company was organizing an independent review of the committee decisions. Participants were told that they would be part of a screening process by which they would have to flag decisions they suspect might be biased. Afterwards, they had to respond to a third and final

comprehension check. In the evaluation phase, the cases were no longer anonymized, but instead showed a portrait picture and the name of each employee in question along with the committee decision (see Figure 1, right panel). Participants could enter their response (biased decision: yes, no) once 1.5 seconds had passed.

After the evaluation phase, participants answered ten self-report questions about their impression of the committee decisions. First, they were asked how often they thought women (men) who did not deserve a pay raise were granted one and vice versa. They could respond on a 7-point Likert scale ("Never", "Very rarely", "Rarely", "Sometimes", "Often", "Very often", "Always"). Next, they judged to what extent the panel decisions exhibited gender bias and reversed gender bias on a 7-point Likert scale ("Not at all" to "Very much"). And finally, they were asked to what extent the panel discriminated against women, discriminated against men and showed favoritism towards women, and showed favoritism towards men, on the same Likert scale as for the previous questions. The descriptive results for the self-report questions are reported in the Online Appendix since they were not part of the confirmatory analyses for Experiment 1. Afterwards, they were asked to briefly describe their strategy in judging the pay raise decisions and to answer demographic questions about their age, gender, highest level of education and ethnicity. Finally, they were asked whether they answered all questions seriously and attentively so that their data could be used for analysis and were given the opportunity to leave a comment. They were debriefed and redirected to Prolific where they received credit for their participation.

Materials

In all experiments, we used portrait pictures showing each one of 128 male- and 128 female-presenting faces that were generated by an adversarial generative network. The portraits were selected based on a validation study in which we collected ratings for the portraits regarding their age, attractiveness, and typicality for a White and Black person

and man and woman.[7] We used a simulated annealing algorithm to select a set of images in which all images were unambiguously categorized in terms of their assigned ethnicity and gender and the subset of men and women were perceived similarly with respect to the mean age and attractiveness. A detailed description, data and analysis scripts of the validation study are provided in the OSF repository. The employees' names were generated from 256 US-American surnames and 128 female and male first names, respectively, commonly used for White US-American citizens. The surnames, first names and portraits were randomly combined without replacement (i.e., no first name, surname or portrait was repeated) for each participant separately and randomly assigned to the cases.

Criteria scores were drawn for each participant separately from a multivariate normal distribution with a mean of 3.5 and moderate correlations between the criteria. If the mean of the scores was larger than 3.5, the correct, unbiased committee decision was set to grant and otherwise deny a pay raise. Consequently, denying a pay raise for an employee with mean score larger than 3.5 would reflect a biased committee decision. For the evaluation phase, only scores from between the 25th and the 75th quantile of the distribution (i.e., scores close to the mean) were used because the cases were supposed to be close calls and as to not make the task too easy for the participants. For the practice phase, scores were not constrained in this manner.

Modeling Approach and Data Analysis

We based our analyses on an equal-variance Gaussian Signal Detection Theory (SDT) model, meaning that the signal (biased cases) and noise (unbiased cases) evidence are assumed to follow Gaussian distributions with different means μ_{signal} and μ_{noise} and the same variance σ^2_{signal} and σ^2_{noise}. In this context, the evidence continuum represents subjective evidence for cues to discrimination in a given situation. A response criterion λ is

[7] The validation study additionally comprised portraits of Black male-presenting and female-presenting people and was used to select the stimuli for this project on gender discrimination and a future project concerned with racial discrimination.

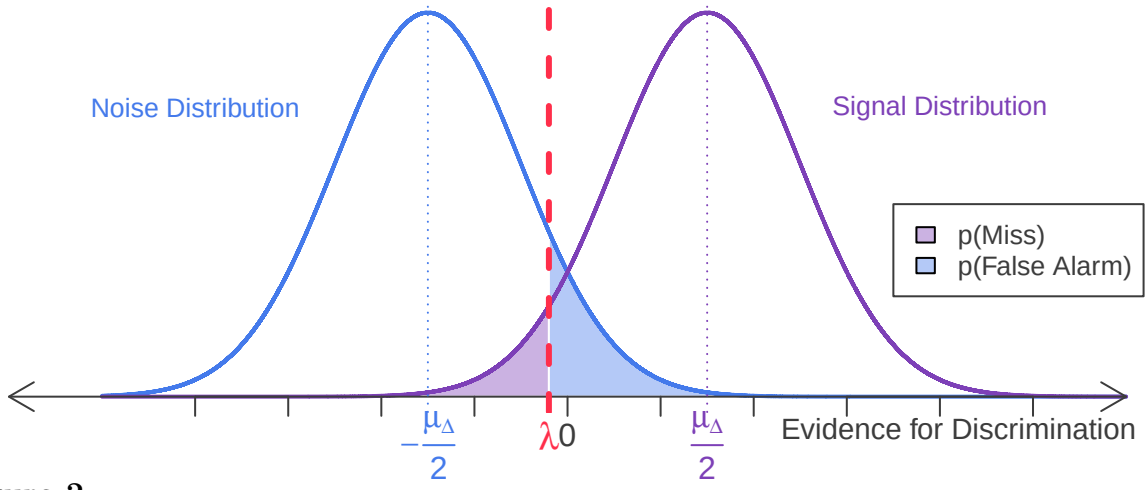

Figure 2
Illustration of the SDT Model and Parametrization

placed on that evidence continuum as a decision threshold: If the evidence surpasses the response criterion, a situation is judged to be "biased", and "unbiased" otherwise. The difference between the means of the signal and noise distributions $\mu_\Delta = \mu_{signal} - \mu_{noise}$ is a measure of sensitivity. As μ_Δ increases, the overlap between the two distributions becomes smaller and the number of both misses (i.e., biased cases judged to be unbiased) and false alarms (i.e., unbiased cases judged to be biased) decreases. The criterion λ, on the other hand, constitutes a measure of response bias: As the response criterion increases, the necessary evidence for a "biased" response increases and the response behavior becomes more *conservative*, meaning that the false alarm rate decreases, whereas the miss rate increases. A higher, more conservative response criterion implies that, overall, fewer biased as well as unbiased cases are judged as being biased. The opposite is true when the criterion decreases; a smaller, more liberal response criterion implies that overall, more cases are judged as biased. Without loss of generality, we set the means of the signal and noise distribution to $\mu_{signal} = \mu_\Delta/2$ and $\mu_{noise} = -\mu_\Delta/2$ and their variance to $\sigma^2_{noise} = \sigma^2_{signal} = 1$. With this parametrization, the intersection of the signal and noise distribution is at 0 and accordingly, a positive response bias indicates a tendency to respond "noise" and a negative response bias indicates a tendency to respond "signal" (see Figure 2).

For all analyses, we estimated SDT parameters and tested our hypotheses jointly using hierarchical SDT models estimated with maximum likelihood in conjunction with likelihood ratio test (for preregistered tests, reported as χ^2 statistics) and Wald (for additional analyses, reported as z statistics) tests on the population-level means (see Rouder & Lu, 2005 for an introduction to hierarchical models). We assumed that participants' individual response bias and sensitivity parameters vary according to Gaussian population distributions characterized by population-level mean (representing mean sensitivity and response bias) and variance parameters (representing the variability between participants in sensitivity and response bias, similar to random intercepts in linear mixed models, see Singmann & Kellen, 2019).[8] We modeled the effects of our experimental manipulations on sensitivity and response bias (e.g., the difference in response bias between decisions where a pay raise was granted and denied) as additive shifts of the respective population mean parameter (i.e., the experimental factors were modeled as a generalized linear model). Given that committee decision and employee gender were perfectly confounded in biased cases in most of our experiments (with the exception of one condition in Experiment 2), we could only estimate the effect of either committee decision or employee gender on sensitivity. We report the results of a model variant including employee gender throughout this manuscript except for analyses directly testing the effect of the committee decision on sensitivity, but an alternative model including the committee decision instead produced similar results and the same substantive conclusions. For factors that were manipulated within participants (i.e., the committee decision, employee gender and their interaction), these additive shifts were in principle also allowed to vary between participants (accounting for differences between participants in the effects of our experimental manipulations, similar to random slopes in linear mixed models). For our preregistered hypothesis tests, we implemented a backward selection such that we started with a full model including all justified by-participant random

[8] This framework can also model by-stimulus variability; however, such models seldomly converged for our data and if they did, the results showed little to no by-stimulus variability in sensitivity and response bias. Therefore, we only included by-participant random effects in our models.

intercepts and slopes (i.e., random intercepts for sensitivity and response bias and random slopes for the effects of the committee decision, employee gender and their interaction on response bias, the effect of the employee gender on sensitivity, and correlations between all random effects) and sequentially removed random effects from the model until it converged. For our hypothesis tests, we tested whether the corresponding fixed effect differs significantly from zero using one-tailed (for directional preregistered hypotheses) and two-tailed likelihood ratio tests. More specifically, we tested whether a model including all population-level and random effects (i.e., all random effects as determined by the backward selection described above) differs significantly from a reduced model containing the same parameters except for the to-be-tested population-level mean (i.e., type III sums of squares tests). For exploratory analyses that were based on different models as preregistered, we included by-participant random intercepts for sensitivity and response bias, and a random slope for the effect of the committee decision on response bias.

Results

After the backward selection procedure, our final model for the preregistered analyses included by-participant variance parameters for mean sensitivity and response bias, for the effect of the committee decision, the employees' gender, and their interaction on response bias, and for the effect of the employee gender on sensitivity.

RQ 1: Base Rates and Cultural Stereotypes

Participants' response bias depended on the interaction between committee decision and employee gender, as indicated by the significant *committee decision* × *employee gender* interaction ($\hat{\beta} = 0.05$, 95% CI $[0.03, 0.07]$, $\chi^2(1) = 27.10$, $p_{\text{two-tailed}} < .001$). As Figure 3 shows, participants' response bias for "granted" and "denied" decisions depended on the gender of the employee in question: As hypothesized, participants were more liberal to judge a decision as biased when a pay raise was granted to a male employee or a pay raise was denied to a female employee than in the remaining cases.

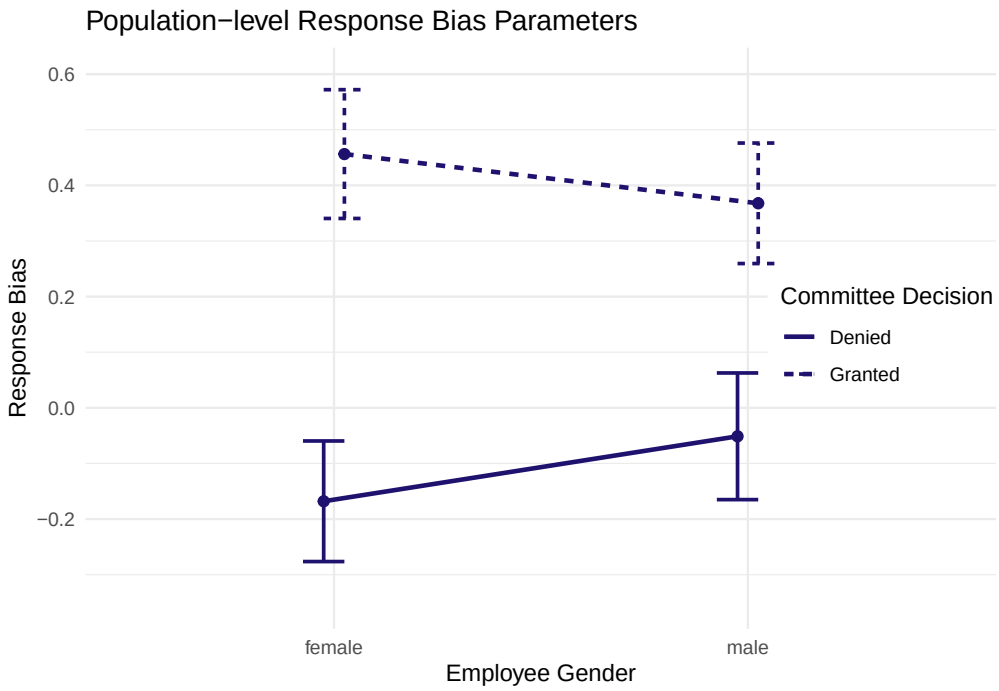

Figure 3

Committee Decision × Employee Gender Interaction. Error bars show 95 % Confidence Intervals (model-based).

RQ 2: Harm

Our analysis revealed a significant effect of the committee decision on participants' response bias: As hypothesized, participants' response bias was smaller for cases in which a pay raise was granted than when a raise was denied ($\lambda_{granted}$ = 0.43, λ_{denied} = -0.12, $\hat{\beta} = 0.27$, 95% CI $[0.17, 0.38]$, $\chi^2(1) = 25.25$, $p_{\text{two-tailed}} < .001$) implying that participants were more liberal (i.e., required less subjective evidence) to judge a decision as biased when a pay raise was denied rather than granted (see Figure 3). The corresponding sensitivity effect was not significantly different from zero ($\mu_{\Delta granted} = 1.65$, $\mu_{\Delta denied} = 1.69$, $\hat{\beta} = -0.02$, 95% CI $[-0.04, 0.00]$, $\chi^2(1) = 1.07$, $p_{\text{two-tailed}} = .302$). Thus, our results are in line with previous findings on the effect of harm and differences between attributions to favoring and disfavoring discrimination (Simon et al., 2019; Swim et al., 2003; York, 1989). Going beyond previous findings, our results indicate that the effect of harm on discrimination judgments reflects response bias rather than a sensitivity effect (cf. Barrett & Swim, 1998).

RQ 3: Gender Differences

Male and female participants did not differ significantly in their response bias ($\lambda_{male} = 0.17$, $\lambda_{female} = 0.14$, $\hat{\beta} = -0.02$, 95% CI $[-0.05, 0.01]$, $\chi^2(1) = 1.16$, $p_{\text{two-tailed}} = .280$). Our analyses did reveal a significant sensitivity effect; however, contrary to our theoretical arguments above, female participants actually exhibited *lower* sensitivity than male participants ($\mu_{\Delta male} = 1.78$, $\mu_{\Delta female} = 1.55$, $\hat{\beta} = -0.11$, 95% CI $[-0.15, -0.07]$, $\chi^2(1) = 6.62$, $p_{\text{two-tailed}} = .010$). A variety of factors may contribute to this effect. Male participants might be more cautious in their judgments compared to female participants: As argued in the Introduction, strong normative pressure against discrimination exists (e.g., Barreto & Ellemers, 2015) and male participants as members of a privileged group might be especially careful in their judgments in order to avoid appearing biased. Higher caution would result in a higher sensitivity while simultaneously increasing the response times of male participants. Thus, this apparent gender difference in sensitivity would reflect differences in the speed-accuracy trade-off (c.f. Fitts, 1966) between male and female participants, with male participants emphasizing accuracy more strongly than female participants. Contrary to this explanation, however, response times of male participants' correct responses ($M_{RT}^{Male} = 4.84$ seconds) were descriptively *faster* than the response times of female participants ($M_{RT}^{Female} = 5.00$ seconds), $F(1, 245) = 3.58$, $p = .059$.[9] Thus, the observed gender difference in sensitivity is unlikely to reflect such differences in the speed-accuracy trade-off.

Another explanation for gender differences in sensitivity invokes a stereotype-threat-like mechanism that impairs the sensitivity of female participants: Threat (such as, for instance, stereotype threat) may decrease performance and in our design, women were more threatened than men by the implemented direction of gender discrimination as well as more threatened by cultural stereotypes on gender bias in the context of pay raise decisions in general. Alternatively, this sensitivity effect may not reflect

[9] For this analysis, we log-transformed latencies and excluded trials with latencies that were extreme values when applying the Tukey criterion to each participants' latencies individually.

anything related to the context of discrimination induced in the BDT: Given that *participant gender* is not an experimentally manipulated factor (i.e., it cannot be assigned randomly), we cannot rule out the possibility that this sensitivity effect reflects an underlying difference between the specific men and women in our participant pool. We will test these explanations in Experiments 2 and 3.

RQ 4: Intergroup Effects

Finally, regarding RQ 4, our results were somewhat inconclusive regarding the hypothesized relative ingroup-favoring response bias: The corresponding *participant gender* × *employee gender* × *committee decision* interaction on response bias was significant in a one-tailed, but not in a two-tailed test ($\hat{\beta} = 0.02$, 95% CI $[0.00, 0.04]$, $\chi^2(1) = 2.88$, $p_{\text{two-tailed}} = .090$). To get a more conclusive result regarding the role of motivated attributions along the lines of intergroup effects in our paradigm, we decided to test this hypothesis again in Experiment 2 with a considerably larger sample that increased statistical power substantially.

Discussion

Taken together, participant's attributions to discrimination in Experiment 1 were aligned with cultural stereotypes about gender discrimination and base rates of discrimination in our design (RQ 1). As hypothesized, these factors affected their response bias rather than their sensitivity. In addition, our results confirmed the effect of harm and differences in attributions to disfavoring versus favoring discrimination and suggest that those factors affect participants' response bias rather than sensitivity (RQ 2). Differences in sensitivity between male and female participants were in the opposite direction from what we expected with male participants having a higher sensitivity than female participants (RQ 3). There were no indications for gender differences in response bias (RQ 4).

In the following experiments, we first consolidate these conclusions through a replication of Experiment 1 (Experiment 2, Stereotypical Bias Condition). We then contrast

the replication results with additional conditions in order to investigate the influence of the direction of gender discrimination (Experiment 2), the discrimination context (Experiment 3), and the nature of participants' decisions (Experiment 4) on our findings to differentiate between different explanations for them (see also Table 1).

Experiment 2: The Direction of Gender Discrimination

The goal of Experiment 2 was to consolidate our previous conclusions through a replication of Experiment 1 and to extend them using additional conditions that allow us to test competing explanations of some of the results. In particular, we explored the role of the direction of gender discrimination in the BDT by contrasting the replication of Experiment 1 (in the following referred to as the *Stereotypical Bias Condition*) with a *Counter-Stereotypical Bias Condition* (in which the committee favors female and disfavors male employees) and a *Symmetrical Bias Condition* (in which the committee favors and disfavors male and female employees in equal proportions). Thereby, we aimed to test competing explanations regarding the *committee decision* × *employee gender* interaction on response bias (RQ 1 - Base Rates and Cultural Stereotypes), the sensitivity difference between male and female participants (RQ 3 - Gender Differences), submit our research question regarding a relative ingroup favoring response bias (RQ 4 - Intergroup Effects) to a higher-powered test and extend our results regarding possible sensitivity effects.

RQ 1 Base Rates and Cultural Stereotypes

We argued that the *committee decision* × *employee gender* interaction on response bias could reflect both cultural stereotypes about gender discrimination in the context of corporate wages as well as the actual base rates of favoring and disfavoring discrimination of male and female employees implemented in our design that participants picked up. Both cultural stereotypes and base rates likely elicit the expectation that the committee will disfavor female and favor male employees which may cause the shift in response bias we observed. The additional conditions we implemented in Experiment 2 allow us to disentangle

the influence of these two factors: If cultural stereotypes about gender discrimination affect participants' response bias, we expect an *employee gender* × *committee decision* interaction across all conditions. If participants also pick up the implemented gender bias of the committee, we expect this interaction to differ between the Conditions, that is, we expect the interaction to be smaller in the Symmetrical Bias and in the Counter-Stereotypical Bias Condition (where it might even reverse in direction).

The Symmetrical Condition additionally allows us to examine a possible sensitivity effect related to cultural stereotypes. As explained above, testing such an effect was not possible in Experiment 1 (and the Stereotypical Bias and Counter-Stereotypical Bias Conditions) because of the perfect confound of employee gender and committee decision in biased trials (i.e., we could not estimate the *employee gender* × *committee decision* interaction on sensitivity). The fully crossed design in the Symmetrical Bias Condition, however, allows us to do so.

RQ 3 (Gender Differences): Threat through Committee Bias

Because we manipulate the direction of gender bias in Experiment 2, the threat of the experimental context for male and female participants changes. In Experiment 1 (and in the Stereotypical Bias Condition of Experiment 2), the committee disfavors female and favors male employees which creates an experimental context that threatens women. The opposite is true for the Counter-Stereotypical Bias Condition, in which the committee favors female and disfavors male employee and accordingly, creates an experimental context that threatens mean. In the Symmetrical Bias Condition the experimental context threatens both groups equally. We argued that increased threat for female participants could have diminished their sensitivity relative to male participants in Experiment 1. If that is the case and the committee's gender bias influences perceived threat, the sensitivity difference between male and female participants should decrease in the Symmetrical Bias Condition and decrease to a stronger degree (or perhaps even reverse) in the Counter-Stereotypical Bias Condition.

RQ 4: Intergroup Effects

As discussed above, the results of Experiment 1 were inconclusive regarding a relative ingroup-favoring response bias. Experiment 2 permits us to test this research question with considerably higher statistical power.

Methods

Experiment 2 was preregistered on the OSF (Stereotypical Bias and Counter-Stereotypical Bias Condition: https://osf.io/qwfe4; Symmetrical Bias Condition: https://osf.io/su9xn).[10]

Sample

1019 participants completed the experiment. $n = 45$ of those participants were excluded ($n = 13$ because their mean sensitivity was close to zero, $n = 28$ because they reported being a different ethnicity than White, $n = 1$ because they were non-binary and $n = 3$ because they stated that their data should not be used for analysis), resulting in a total sample size of $N = 974$, with $n = 245$ participants in the Stereotypical Bias Condition (122 women), $n = 242$ participants in the Counter-Stereotypical Bias Condition (120 women) and $n = 487$ participants in the Symmetrical Bias Condition (246 women). The mean age of the sample was $M = 33.81$ ($SD = 7.40$), ranging from 18 to 49. Our target sample size (120 women and men each in the Stereotypical Bias and Counter-Stereotypical Bias Conditions; 240 men and women each in the Symmetrical Bias Condition) was based on the same power analysis as in Experiment 1. To compensate for the decrease in power due to the smaller number of trials per cell in the Symmetrical Bias Condition (see Section "Design and

[10] Note that we deviated slightly from the preregistered analysis plan: We additionally included the data from the Symmetrical Bias Condition (which was preregistered as a separate experiment) in the analysis of the Stereotypical Bias and Counter-Stereotypical Bias Conditions to be able to compare all three contingency conditions at once. The pattern of results and conclusions are the same as in the preregistered analysis. A summary of the preregistered analysis is available in the Online Appendix; the analysis code is provided in the OSF repository. Furthermore, we preregistered replication hypotheses for the pattern of effects found in Experiment 1 which we do not report and discuss here, but only in the Online Appendix because we replicated all effects, except for the ingroup-favoring response bias which we discuss below.

Procedure"), we doubled our target sample size for this condition.

Design and Procedure

Experiment 2 followed the same design as Experiment 1 with an additional between-participants factor *stimulus contingencies* (Stereotypical Bias vs. Symmetrical Bias vs. Counter-Stereotypical Bias). The Stereotypical Bias Condition implemented the same stimulus contingencies as Experiment 1. In the Counter-Stereotypical Bias Condition, the contingency pattern was reversed, such that the 64 biased trials comprised 32 trials in which female employees were unfairly granted a pay raise and 32 trials in which male employees were unfairly denied a pay raise. In the Symmetrical Bias Condition, 160 out of the 256 cases comprised fair decisions, and the remaining 96 cases comprised unfair decisions. The 96 biased trials comprised 24 trials in which a female employee was unfairly denied a pay raise, 24 trials in which a female employee was unfairly granted a pay raise, 24 trials in which a male employee was unfairly denied a pay raise and 24 trials in which a male employee was unfairly granted a pay raise. Participants were randomly assigned to the Stereotypical Bias or Counter-Stereotypical Bias Condition using a block-randomization procedure to ensure a similar number of participants per condition. Data for the Symmetrical Bias Condition were collected in a separate data-collection effort (see section "The Present Research"). The procedure was similar to Experiment 1 for all conditions[11].

Results

After the backward selection procedure, our final model for the preregistered analyses included by-participant variance parameters for mean sensitivity and response bias and for the effect of the committee decision on response bias.

---

[11] We implemented some minor procedural changes in Experiment 2 compared to Experiment 1: We changed one question in a comprehension check, separated the evaluation phase into four blocks of 64 trials each and slightly rephrased the summary ratings at the end of the experiment (see the Online Appendix for details).

RQ 1: Base Rates versus Cultural Stereotypes

The significant *committee* × *employee gender* interaction replicated Experiment 1 and indicated that participants' response bias was more liberal when a pay raise was granted to a male employee and denied to a female employee than in the remaining cases ($\hat{\beta} = 0.02$, 95% CI $[0.01, 0.03]$, $\chi^2(1) = 18.27$, $p_{\text{one-tailed}} < .001$). This finding suggests that across all conditions, participants' response bias reflected cultural stereotypes about gender discrimination, even when the committee exhibited no bias or even counter-stereotypical gender bias.

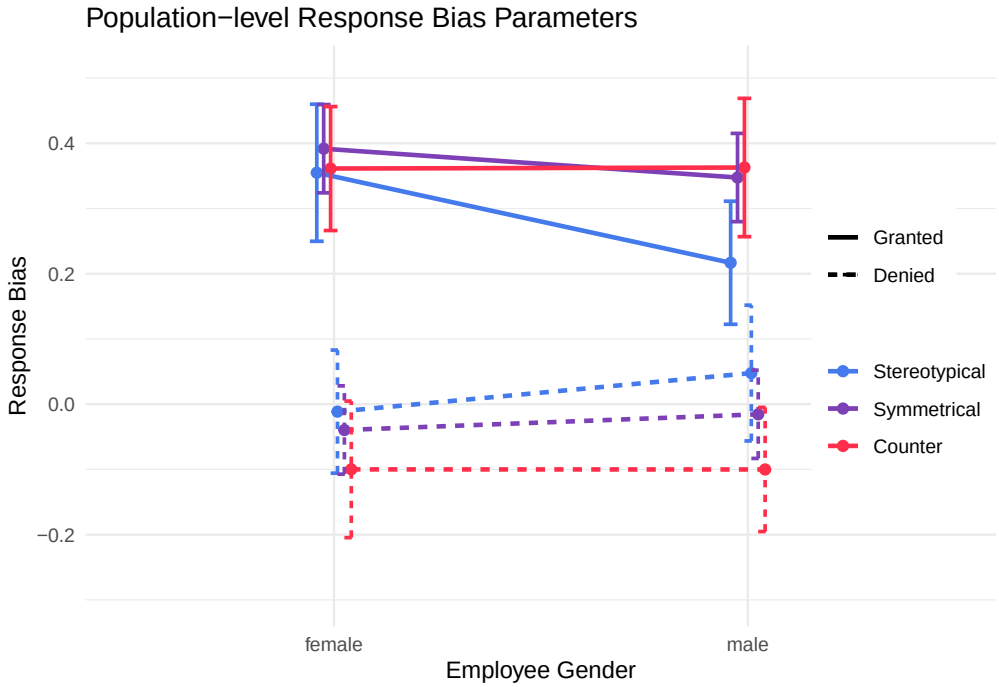

Figure 4

Committee Decision × Employee Gender Interaction in the Stereotypical Bias (abbreviated as "Stereotypical"), Counter-Stereotypical Bias ("Counter") and Symmetrical Bias ("Symmetrical") Condition. Error bars show 95 % Confidence Intervals (model-based).

Additionally, the size (but not the direction) of this interaction differed between the conditions, as Figure 4 makes apparent ($\chi^2(1) = 21.05$, $p_{\text{two-tailed}} < .001$). Contrasts comparing the Stereotypical and Symmetrical Bias Conditions ($\Delta\hat{\beta} = 0.13$, 95% CI $[0.05, 0.21]$, $z = 3.31$, $p < .001$), the Symmetrical and Counter-Stereotypical Bias Conditions

($\Delta\hat{\beta} = 0.07$, 95% CI $[-0.01, 0.15]$, $z = 1.77$, $p = .076$), and the Stereotypical and Counter-Stereotypical Bias Conditions ($\Delta\hat{\beta} = 0.20$, 95% CI $[0.11, 0.29]$, $z = 4.44$, $p < .001$) indicate exactly the expected pattern suggested by the implemented stimulus contingencies. That is, the more cultural stereotypes and stimulus contingencies align, the stronger the interaction.[12]

Participants' ratings of the committee's gender bias show that they explicitly picked up on the implemented contingencies (Stereotypical Bias Condition: $M_{Bias} = 1.00$, Symmetrical Bias Condition: $M_{Bias} = 1.76$, Counter-Stereotypical Bias Condition: $M_{Bias} = 2.37$, $F(2, 971) = 197.11$, $p < .001$; with rating labels 0 = stereotypical gender bias, 2 = no gender bias, 4 = reversed gender bias). When including the mean bias rating in the model, it predicted the size of the *committee* × *employee gender* interaction (stronger perceived gender bias against women was associated with a stronger interaction, $\hat{\beta} = 0.02$, 95% CI $[0.01, 0.03]$, $z = 5.67$, $p < .001$), indicating that participants' perception of the bias of the committee affected their response bias.

Regarding sensitivity, the analysis of the Symmetrical Bias Condition did not provide evidence for a corresponding *employee gender* × *committee decision* interaction ($\hat{\beta} = 0.00$, 95% CI $[-0.01, 0.01]$, $z = 0.24$, $p = .812$). Given that the committee in the Symmetrical Bias Condition neither favors nor disfavors men versus women and thus there are no such stimulus contingencies participants might pick up, this interaction only reflects cultural stereotypes here. Thus, participants' sensitivity seems to be unrelated to such stereotypes and corresponding expectations which did, however, affect response bias and contingency ratings in this condition.

---

[12] A limitation of this study is that the data for the Symmetrical Bias Condition was collected in a separate recruitment effort, and therefore, participants were not strictly randomly assigned to the three conditions. However, given that only considering the Stereotypical Bias and Counter-Stereotypical Bias Conditions (where participants were randomly allocated to one or the other) leads to similar results (see the Online Appendix), our conclusions are not affected by this limitation.

RQ 3: Gender Differences

Threat through Committee Bias. Experiment 2 again revealed larger sensitivity for male than female participants ($\hat{\beta} = -0.13$, 95% CI $[-0.15, -0.11]$, $\chi^2(1) = 35.56$, $p_{\text{one-tailed}} < .001$). This difference was not affected by the contingencies ($\hat{\beta} = 0.00$, 95% CI $[-0.03, 0.03]$, $\chi^2(1) = 0.00$, $p_{\text{one-tailed}} > 0.999$ for the contrast between the Stereotypical Bias and Counter-Stereotypical Bias Conditions, $\hat{\beta} = 0.06$, 95% CI $[0.03, 0.09]$, $\chi^2(1) = 5.20$, $p_{\text{one-tailed}} > 0.999$ for the contrast between the Counter-Stereotypical Bias and Symmetrical Bias Conditions, see Figure 5). Thus, our results clearly contradict the explanation that the direction of gender discrimination impairs performance of participants whose gender that is disfavored in the design through an experimental context threat.[13]

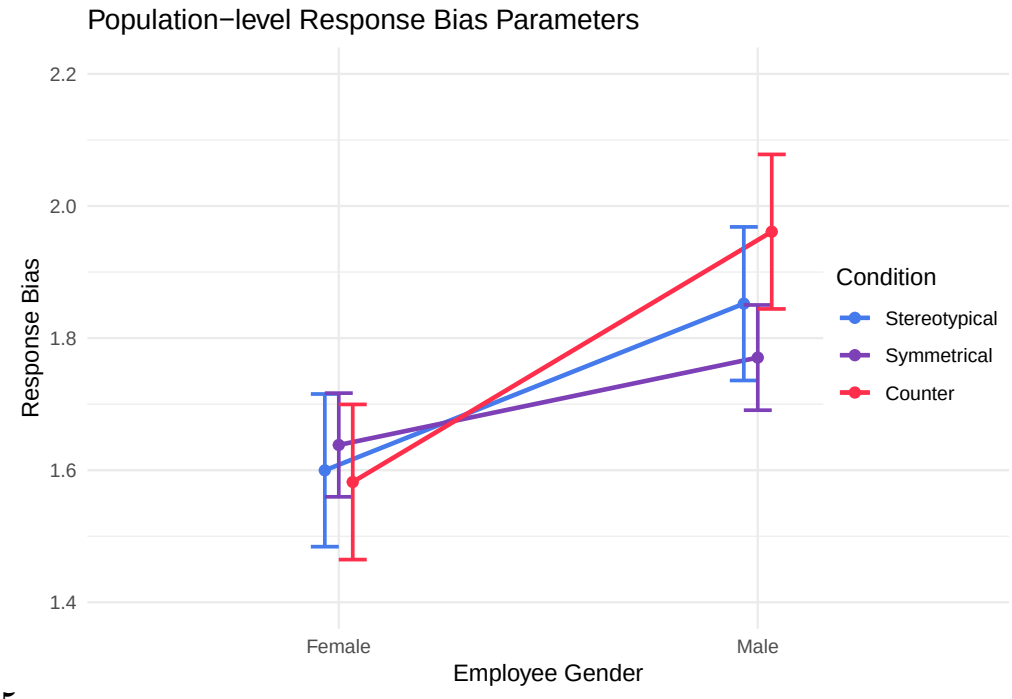

Figure 5

Sensitivity Difference Between Male and Female Participants in the Stereotypical Bias (abbreviated as 'Stereotypical'), Counter-Stereotypical Bias ('Counter') and Symmetrical Bias ('Symmetrical') Condition. Error bars show 95 % Confidence Intervals (model-based).

---

[13] Again, this conclusion is unaffected by the non-random group allocation for the Symmetrical Bias Condition, because the pattern of results is equivalent when only considering the Counter-Stereotypical Bias and Stereotypical Bias Conditions.

Response Bias of Male and Female Participants. Whereas the results of Experiment 1 did not provide evidence for gender differences in response bias, an exploratory analysis of Experiment 2 suggests that female participants were more liberal than male participants in judging the committee decisions as being biased ($\hat{\beta} = -0.03$, 95% CI $[-0.05, -0.01]$, $\chi^2(1) = 8.33$, $p_{\text{two-tailed}} = .004$). This effect was not significantly moderated by the contingency condition ($\hat{\beta} = -0.01$, 95% CI $[-0.04, 0.02]$, $z = -0.73$, $p = .464$ for the contrast between the Stereotypical Bias and Counter-Stereotypical Bias Conditions, $\hat{\beta} = 0.01$, 95% CI $[-0.01, 0.03]$, $z = 0.76$, $p = .446$ for the contrast between the Symmetrical Bias and Counter-Stereotypical Bias Conditions). The most likely explanation for the diverging results is the considerably higher statistical power in Experiment 2. Thus, there seems to be empirical support for differences in response bias between male and female participants, although the effect is rather small.

RQ 4: Relative Ingroup-Favoring Response Bias

The three-way interaction corresponding to the relative ingroup-favoring response bias was neither significant in the preregistered analysis (considering only Stereotypical Bias Condition; $\hat{\beta} = 0.00$, 95% CI $[-0.02, 0.01]$, $\chi^2(1) = 0.08$, $p_{\text{one-tailed}} = .388$), nor in an analysis including all conditions ($\hat{\beta} = 0.00$, 95% CI $[-0.01, 0.01]$, $\chi^2(1) = 0.04$, $p_{\text{two-tailed}} = .843$). Thus, the results of Experiment 2 do not provide evidence for such an ingroup-favoring response bias pattern in the BDT.

RQ 2: Sensitivity Effects of Harm

Our preregistered exploratory analysis of only the Symmetrical Bias Condition revealed that participants' sensitivity was higher for "denied" than for "granted" decisions ($\mu_{\Delta,granted} = 1.70$, $\mu_{\Delta,denied} = 1.77$, $\hat{\beta} = -0.04$, 95% CI $[-0.05, -0.03]$, $\chi^2(1) = 11.01$, $p_{\text{two-tailed}} < .001$), which dovetails with effects of harm on sensitivity that were hypothesized in the literature but not with our results from Experiment 1. Again, the most likely explanation for the diverging results is the difference in statistical power: Although the

Symmetrical Bias Condition had a smaller number of observations per cell than Experiment 1 (24 vs. 32), the sample was twice as large, likely resulting in a higher power to detect smaller effects. In line with this argument, the effect is also significant in the analysis of all three conditions ($\hat{\beta} = -0.03$, 95% CI $[-0.04, -0.03]$, $\chi^2(1) = 13.46$, $p_{\text{two-tailed}} < .001$) and not significantly moderated by the contingency condition ($\hat{\beta} = 0.01$, 95% CI $[0.00, 0.03]$, $z = 1.42$, $p = .154$ for the contrast between the Stereotypical Bias and Counter-Stereotypical Bias Conditions, $\hat{\beta} = -0.01$, 95% CI $[-0.02, 0.00]$, $z = -1.36$, $p = .173$ for the contrast between the Symmetrical Bias and Counter-Stereotypical Bias Conditions). Thus, the results of Experiment 2 suggest that both sensitivity and response bias appear to reflect the harm of the decision outcome or differences in people's perception of favoring and disfavoring discrimination – with a more robust and larger effect on response bias than sensitivity.

Discussion

Taken together, the results of Experiment 2 imply that both cultural stereotypes about gender discrimination and the actual direction of gender discrimination in a specific context affected participants' response bias (RQ 1: Base Rates and Cultural Stereotypes). Participants' sensitivity, however, was not affected by whether or not cultural stereotypes suggested heightened likelihood of bias in judged cases. In addition, there were indications that both response bias and sensitivity are influenced by harm (higher in cases involving harmful, disfavoring discrimination than in cases showing beneficial, favoring discrimination, RQ 2) and that women might have a somewhat more liberal response bias than men (RQ 3: Gender Differences). We suspect that the power in Experiment 1 was too low to detect the effects of harm on sensitivity and gender differences in response bias; however, additional confirmatory evidence is necessary to support these latter conclusions. Our results speak against an effect of the direction of the committee's gender bias on participants' sensitivity and thereby against an explanation in terms of threat for the robust sensitivity difference between male and female participants (RQ 3: Gender Differences). However, these results

certainly do not rule out threat as an explanation altogether: Women are in general more threatened by the subject of gender bias than men, and thus a more general threat caused by the discrimination context (which was still present in all conditions in Experiment 2) might diminish the sensitivity of female participants independent of the implemented direction of bias in our design. Alternatively, as explained above, because *participant gender* is not an experimental factor, this effect may reflect differential selection mechanisms by which participants come to participate in the study as a function of participant gender so that possible differences in sensitivity might reflect selection bias unrelated to discrimination. We tested these explanations in Experiment 3.

Experiment 3: The Discrimination Context

In Experiment 3, we explored the role of the discrimination context on participants' judgments in the BDT. To do so, we implemented a condition aimed at eliminating the context associated with likely discrimination, the *Placement Condition*, in which participants had to judge placement decisions instead of pay raise decisions.[14] The placement decisions were concerned with the assignment of employees to one of two departments with similar prestige and equivalent wages. Like the pay raise decisions, placement decisions were based on ratings which in this context indicated the employee's relative fit for the two departments. Except for these crucial differences, the Placement Condition was identical to the BDT, as implemented in Experiments 1 and 2.

RQ 3: Gender Differences

The Placement Condition allowed us to further interrogate our explanations for the difference in sensitivity between male and female participants: If stereotype threat (Nguyen & Ryan, 2008) induced by the discrimination context in the BDT or another variable linked to the discrimination context in the task diminishes the sensitivity of female participants, the

[14] The Placement Condition is contrasted with the Stereotypical Bias Condition described in Experiment 2. Note that participants had been randomly assigned to these conditions during data collection (see section "The Present Research").

sensitivity difference between male and female participants should be smaller (or even completely eliminated) in the Placement Condition. If, on the other hand, this difference reflects variables that are unrelated to the discrimination context, such as selection bias, we expect a similar difference in both conditions.

RQ 1: Base Rates and Cultural Stereotypes

Additionally, the Placement Condition positions us to strengthen our conclusions regarding the influence of cultural stereotypes and base rates on response bias and to examine their dependence on the discrimination context: The Placement Condition presents participants with the same base rates as in the BDT implemented in Experiment 1 and the Stereotypical and Counter-Stereotypical Bias Conditions in Experiment 2, but cultural stereotypes about gender bias should not affect judgments in the Placement Condition because the decisions regard evaluatively equivalent placements instead of pay raises. Thus, we expect the *committee decision* × *employee gender* interaction to be reduced in the Placement Condition.

Methods

The preregistration for Experiments 3 and 4 can be found at https://osf.io/e7fw4.[15]

Sample

246 participants completed the Placement Condition. 6 of those participants were excluded ($n = 4$ because they reported a different ethnicity than White, and $n = 2$ because they stated that their data should not be used for analysis), resulting in a total sample size of $N = 240$ (120 women). The mean age of the sample was $M = 35.78$ ($SD = 7.82$), ranging from 18 to 49.

---

[15] Experiments 3 and 4 were preregistered as one experiment but we present them and the corresponding preregistered analyses here separately for a more coherent flow.

Design

The design of Experiment 3 was again similar to Experiment 1 with an additional factor *Study Condition* (Stereotypical Bias Condition vs. Placement Condition). We implemented similar stimulus contingencies in the Placement Condition as in the Stereotypical Bias Condition, such that in cases with erroneous decisions, female employees were always misplaced to one and male employees to the other department (which department men and women were misplaced to was randomly determined for every participant). For the analysis, the committee decisions in the Placement Condition were internally recoded such that the contingencies were equivalent to the Stereotypical Bias Condition.

Procedure

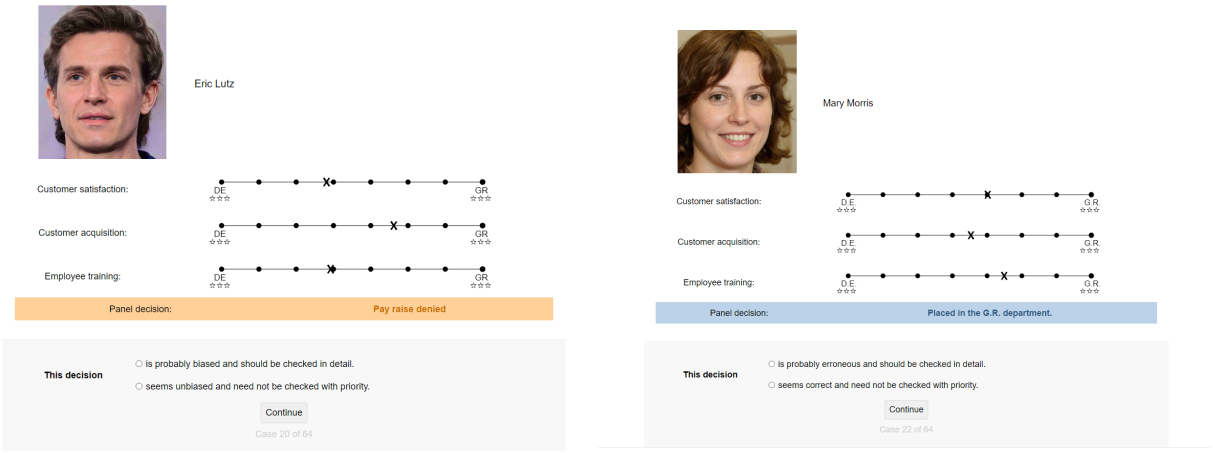

Figure 6

Example trials of the evaluation phase in the Stereotypical Bias (left) and Placement Condition (right).

The procedure of the Placement Condition was aligned as closely as possible to the Stereotypical Bias Condition. Participants were instructed that they would have to evaluate a series of decisions regarding the *placement* of trainees in one of two departments. They were told that the departments had a similar reputation and that the employees were paid the same wage, but that the employees might have different aptitudes for the two

departments and placing them in the correct department would lead to a better performance and higher job satisfaction. As in the Stereotypical Bias Condition, placement decisions would be made based on scores on three criteria which in the Placement Condition indicated the relative fit to the two departments named "DE" and "GR" (see Figure 6). After the practice phase, participants were informed about complaints about *erroneous* placement decisions and that their task was to flag decisions that appeared erroneous to them. The comprehension checks were adapted to the context of the placement decisions but were otherwise identical to the Stereotypical Bias Condition. One comprehension check comprised a question about the reputation and wage of the two departments to ensure that participants understood that there was no inherent difference in value related to either pay or prestige between the two departments. After the evaluation phase, participants answered the same self-report questions as the participants in the Stereotypical Bias Condition.

Results

After the backward selection procedure, our final model for the preregistered analyses included by-participant variance parameters for mean sensitivity and response bias and for the effect of the committee decision on response bias.

Manipulation Check

To ensure that we successfully reduced the perception of discrimination in the Placement Condition, we compared participants' responses to the self-report questions about the committee decisions between the study conditions. Compared to participants in the Placement Condition, participants in the Stereotypical Bias Condition perceived female employees to be hurt more than male employees ($\Delta M = -0.84$, 95% CI $[-0.96, -0.72]$, $t(418.29) = -13.61$, $p < .001$ on a scale from 0 to 4) and favored less than male employees ($\Delta M = 0.96$, 95% CI $[0.83, 1.09]$, $t(446.14) = 14.70$, $p < .001$ on a scale from 0 to 4). Similarly, participants in the Stereotypical Bias Condition indicated a stronger gender bias of the committee ($\Delta M = 0.84$, 95% CI $[0.71, 0.96]$, $t(465.02) = 13.12$, $p < .001$ on a scale from

0 to 4) and that the panel discriminated against women more ($\Delta M = 0.81$, 95% CI $[0.69, 0.93]$, $t(470.00) = 13.40$, $p < .001$ on a scale from 0 to 4) compared to the Placement Condition.

Furthermore, participants seemed to perceive misplacements to both departments as similarly harmful – neither as favoring nor disfavoring – as indicated by the non-significant difference in response bias between the two types of decisions ($\Delta = 0.00$, 95% CI $[-0.15, 0.15]$, $z = -0.05$, $p = .964$) in the Placement Condition. This difference was significantly smaller in the Placement Condition than in the Stereotypical Bias Condition ($\hat{\beta} = -0.12$, 95% CI $[-0.22, -0.01]$, $\chi^2(1) = 4.96$, $p_{\text{one-tailed}} = .013$).

RQ 3 (Gender Differences): Discrimination Context Threat

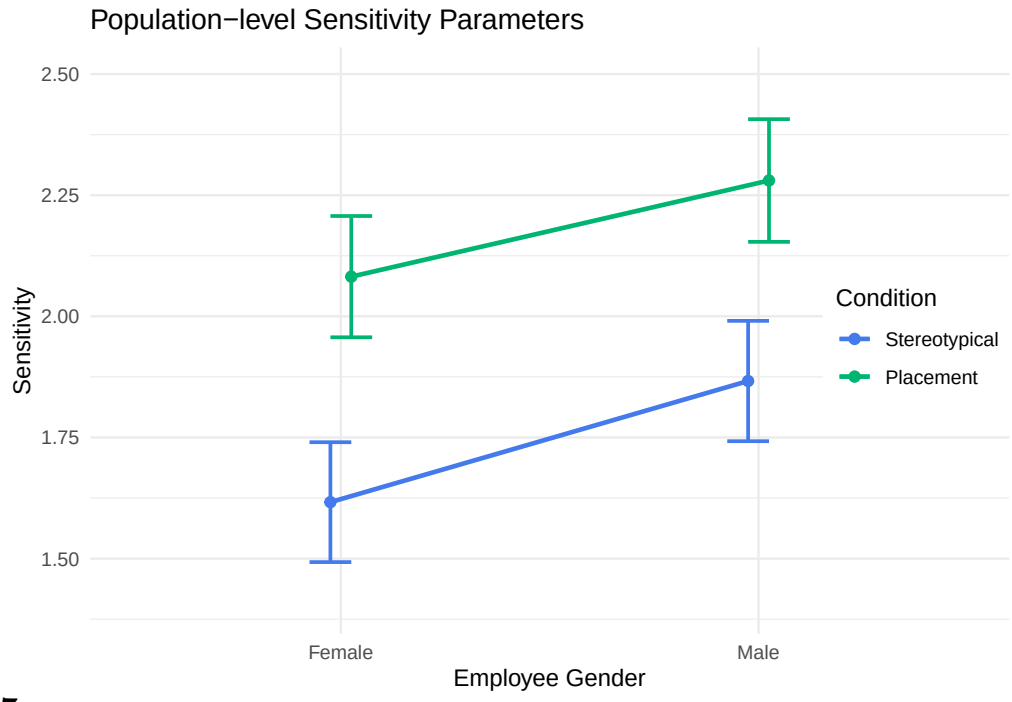

Figure 7

Sensitivity Difference Between Male and Female Participants in the Stereotypical Bias and Placement Condition. Error bars show 95 % Confidence Intervals (model-based).

Across both conditions, male participants showed greater sensitivity than female participants ($\hat{\beta} = -0.06$, 95% CI $[-0.11, -0.02]$, $z = -2.80$, $p = .005$) The sensitivity difference between male and female participants was not significantly different between the

Stereotypical Bias and the Placement Condition ($\hat{\beta} = 0.03$, 95% CI $[-0.04, 0.09]$, $\chi^2(1) = 0.16$, $p_{\text{two-tailed}} = .687$, see Figure 7). This result contradicts the explanation relating the sensitivity difference to stereotype threat occurring for women in the Stereotypical Bias Condition and other explanations related to the discrimination context in general. Thus, our results do not allow us to attribute the sensitivity difference to threat caused by the discrimination context. At this point, the most likely explanation so far is a selection effect with some difference in how men and women come to participate in the experiment causing the difference in sensitivity between men and women.

RQ 1 (Base Rates and Cultural Stereotypes): Discrimination Context

Figure 8

Committee Decision × Employee Gender Interaction in the Stereotypical Bias and Placement Condition. Error bars show 95 % Confidence Intervals (model-based).

The *committee decision* × *employee gender* interaction was significantly smaller in the Placement Condition than in the Stereotypical Bias Condition ($\hat{\beta} = -0.04$, 95% CI $[-0.06, -0.02]$, $\chi^2(1) = 11.01$, $p_{\text{one-tailed}} < .001$) and did not differ significantly from zero in the former condition ($\Delta = -0.01$, 95% CI $[-0.03, 0.00]$, $z = -1.41$, $p = .157$, see Figure 8).

This finding supports our previous conclusions regarding the effect of cultural stereotypes and base rates. In fact, stimulus contingencies did not appear to influence participants' response bias *at all* in the Placement Condition. Participants' judgments about the frequency of different committee decisions provide an explanation for this: In the Stereotypical Bias Condition, participants detected the contingencies and explicitly reported the frequency of cases as reflected in the experimental manipulation (i.e., women being denied a pay raise they would have deserved and men being granted a raise even though they did not deserve one) to be significantly higher than the frequency of the remaining cases (i.e., men being denied a pay raise they would have deserved and women being granted a raise even though they did not deserve one; $M_{Stereotypiycal} = 2.62$, 95% CI $[2.35, 2.89]$, $t(244) = 19.00$, $p < .001$). In the Placement Condition, this explicit contingency learning was much weaker ($M_{Placement} = 0.20$; $\Delta M = 2.42$, 95% CI $[2.11, 2.73]$, $t(393.09) = 15.16$, $p < .001$ for the difference between the Placement and the Stereotypical Bias Condition), in line with and probably accounting for the reduced effect of stimulus contingencies on response bias in this condition. Thus, the discrimination context did not only influence participants' response bias through cultural stereotypes, but also bolstered participants' contingency learning regarding the committee decisions and the employees' gender.

Discussion

Taken together, Experiment 3 suggests that contingency learning is stronger in contexts associated with discrimination, in which the social categories on which discrimination is based are perceived to be more vital. Contingency learning in turn correlated with response bias. Our results, however, contradict the explanation relating the sensitivity difference between men and women to the discrimination context through increased threat for female participants in the Stereotypical Bias Condition relative to the Placement Condition.

Experiment 4: Decision Processes

Finally, we aimed to probe even deeper into participants‘ decision processes leading to attributions to discrimination. As pointed out in the Introduction, attributions to discrimination regularly involve a comparison between an assessment of what constitutes fair treatment and the observed treatment in determining whether an individual was treated unjustly. Accordingly, attributions to discrimination based on such comparisons are shaped by two components: (1) judgments of what constitutes fair treatment (referred to hereafter as *standard assessment*) and (2) comparisons of that standard with actual treatment. Effects of different variables (e.g., characteristics of victims, actors, or situational factors) on attributions can therefore reflect effects on standard assessments and/or effects on comparisons, as well as on further attributional processes following from such comparisons (e.g., an attribution to the victim's group membership). Interestingly, standard assessment can in principle occur completely independently of actual treatment and thus independently of the act or decision that is to be judged. Effects of third variables may thereby reflect effects on standard assessment independently of the observed decision rather than effects on attributional processes relating to the decision as such -– a possibility that we believe has not received adequate research attention so far.

To illustrate this thought in some more detail, consider two effects consistently observed in the previous experiments: (1) the difference in response bias between "denied" and "granted" decisions and (2) the interaction between committee decision and employee gender in response bias. The first effect would traditionally be interpreted as an effect of harm caused by the observed committee decision on the observers' inclination to attribute committee decisions to discrimination. Alternatively, this effect may reflect an overall bias in standard assessment to consider employees as generally more deserving of pay raises (i.e., a beneficial treatment) rather than not deserving them (i.e., deserving a more harmful treatment – the denial of a pay raise), independently of any actual committee decision. Such a bias in standard assessment would suffice to cause observers to consider committee

decisions unfair more frequently when a pay raise is denied than when one is granted because the committee decisions are simply more likely to match the beneficial standard when they are themselves beneficial (i.e., when a pay raise is granted). Thereby, such a bias in standard assessment can account for the observed response bias effect in the absence of any effect of observed harm on the attributional processes themselves.

As another example, consider (2) the interaction between committee decision and employee gender such that observers have a stronger tendency to consider committee decisions to be biased when a woman is denied or a man is granted a pay raise than in the remaining cases. This interaction would traditionally be explained in terms of cultural stereotypes about likely bias in committee decisions and, in the case of our experiments, in terms of the direction of gender bias implemented via the stimulus contingencies. According to this explanation, participants would use the actual committee decision in conjunction with the employee's gender to adjust their inclination to judge the committee decision as being biased. Alternatively, participants may hold men and women to different standards such that they are more willing to consider women as deserving a pay raise than men -– a pro-women bias in standard assessment that might occur independently of the actual committee decision, but could account for the observed interaction in attributions to bias.

Experiment 4 capitalizes on a comparison of the current paradigm in the Stereotypical Bias Condition and a version of the judgment bias task (Axt et al., 2018) to disentangle these possibilities. For this purpose, we contrasted the Stereotypical Bias Condition with a Decision Condition in which -– similar to the judgment bias task — participants had to make the pay raise decisions themselves, in the absence of any committee decision, instead of judging given pay raise decisions. The Decision Condition thereby allows us to establish a baseline of what participants consider an appropriate treatment, independently of any committee decision. This setup allows us to determine the extent to which the effects that we observe reflect effects on standard assessment independently of committee decision and the extent to which they reflect effects on attributional processes

that take actual committee decisions into account.

Methods

The preregistration for Experiments 3 and 4 can be found at https://osf.io/e7fw4.[16]

Sample

246 participants completed the Decision Condition. 3 of those participants were excluded because they reported being a different ethnicity as White, resulting in a total sample size of $N = 243$ (122 women). The mean age of the sample was $M = 35.14$ ($SD = 7.93$), ranging from 18 to 49.

Design

Again, the design was similar to Experiment 1 with an additional between-subjects factor *Study Condition* (Stereotypical Bias Condition vs. Decision Condition). Although the Decision Condition did not present any committee decisions, we tacitly assigned a committee decision to each case with the same stimulus contingencies as in the Stereotypical Bias Condition to be able to compare the conditions directly.

Procedure

The procedure was again similar to the BDT with the exception of the evaluation phase, in which participants' task was to continue to make pay raise decisions. Before the evaluation phase, participants in the Decision Condition received no instruction about biased decisions and instead were told that they would continue with their task from the practice phase (i.e., *making* the pay raise decisions) but that they would not see a typical decision after they had made their decision anymore. The third comprehension check was adapted accordingly. There were no self-report questions at the end of the experiment in the Decision Condition.

[16] Experiments 3 and 4 were preregistered as one experiment but we present them and the corresponding preregistered analyses here separately for a more coherent flow (see also section "The Present Research").

Data Analysis

In the Decision Condition, participants' responses were coded as "biased" and "unbiased" depending on whether they agreed with the internally assigned committee decision. This recoding allowed us to analyze the Decision and the Stereotypical Bias Condition in the same model. Note however that the model parameters must be interpreted somewhat differently than in the Stereotypical Bias Condition; for instance, as explained above, an effect of the factor *committee decision* in the Decision Condition would reflect a bias to grant or deny pay raises (or vice versa).

Results

After the backward selection procedure, our final model for the preregistered analyses included by-participant variance parameters for mean sensitivity and response bias and for the effect of the committee decision on response bias.

Pro-Women Bias in Standard Assessment versus Gender Bias Relating to Committee Decisions

As already mentioned in Experiment 3, the *employee gender* × *committee decision* interaction was significant. It was significantly smaller in the Decision Condition than in the Stereotypical Bias Condition ($\hat{\beta} = -0.04$, 95% CI $[-0.06, -0.02]$, $\chi^2(1) = 12.58$, $p_{\text{two-tailed}} < .001$) and not significantly different from zero in the former ($\Delta = -0.01$, 95% CI $[-0.03, 0.00]$, $z = -1.58$, $p = .114$, see Figure 9). This result suggests that the interaction found in the Stereotypical Bias Condition does not reflect a pro-women bias in standard assessment – a possibility rendered likely by the occurrence of such bias in the judgment bias task (Axt et al., 2018). Instead, this interaction appears to be mediated by the presence of the committee decision, and likely reflects differences in perceived biasedness of those decisions as a function of cultural stereotypes and stimulus contingencies, rather than perceived deservingness,

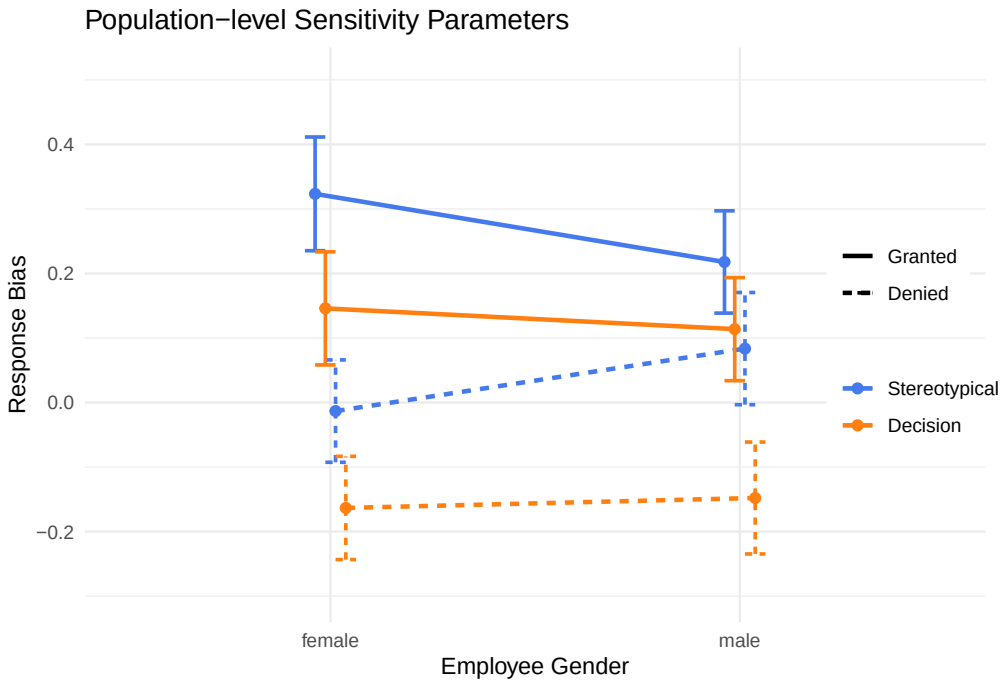

Figure 9

Committee Decision × Employee Gender Interaction in the Stereotypical Bias and Decision Condition. Error bars show 95 % Confidence Intervals (model-based).

Bias to Grant Pay Raises versus Harm of Committee Decision

The effect of *committee decision* in the Decision Condition was similar to (and not significantly different from) the Stereotypical Bias Condition ($\hat{\beta} = 0.03$, 95% CI $[-0.08, 0.13]$, $\chi^2(1) = 0.22$, $p_{\text{two-tailed}} = .637$, see Figure 9). This pattern of results suggests that the corresponding effect in the Stereotypical Bias Condition (i.e., the more liberal response bias for "denied" than "granted" decisions) reflects an effect on standard assessment such that participants are inclined to grant rather than deny pay raises. This result thus contradicts the notion that the effect of committee decision on response bias reflects differences in the perception of disfavoring versus favoring discrimination because participants exhibit the same bias when that decision is not shown. However, the effect does not necessarily contradict the effect of harm because participants likely consider the consequences or harm of the two decisions when making them themselves. Thus, harm might already affect both participants' judgments of deservingness as a precursor to their judgments about the

biasedness of the committee decisions.

Sensitivity of Men and Women

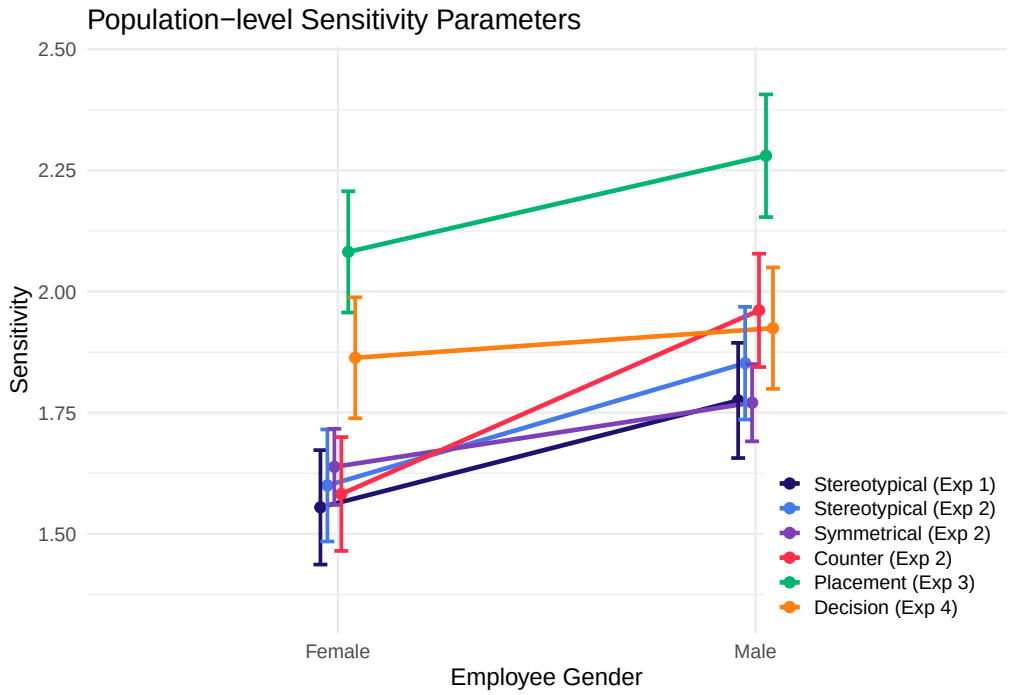

Figure 10

Sensitivity Difference Between Male and Female Participants across all experiments. Error bars show 95 % Confidence Intervals (model-based).

The sensitivity difference between male and female participants was not significantly different between the Stereotypical Bias and the Decision Condition in our preregistered hypothesis test ($\hat{\beta} = 0.09$, 95% CI $[0.03, 0.16]$, $\chi^2(1) = 2.20$, $p_{\text{two-tailed}} = .138$). However, Figure 10 illustrates that while there is a difference in sensitivity in all conditions employing an attribution task (i.e., the BDT and the Placement Condition from Experiment 3), this difference is close to zero (and in fact not significantly different from zero, $\Delta = -0.06$, 95% CI $[-0.24, 0.12]$, $z = -0.68$, $p = .498$) in the Decision Condition. This result is difficult to interpret because there was no significant difference between the Stereotypical Bias Condition and the Decision Condition in our preregistered hypothesis test. Still, if the result (i.e., the absence of a sensitivity difference in the Decision Condition) proves to be reliable in future studies, it would contradict a simple account in terms of selection bias because it is

difficult to conceive of a difference in our populations of men and women that results in men being more accurate in judging pay raise decisions and placement decisions, but not when judging deservingness. Why exactly men and women might differ in their sensitivity in attribution tasks (i.e., the BDT and the Placement Condition) but not in the corresponding judgment task, however, remains an open question.

Discussion

The results of Experiment 4 suggest that the interaction between committee decision and employee gender reflects participants' *beliefs about the committee's gender bias* (in terms of both cultural stereotypes about gender bias and base rates) rather than a pro-women bias in standard assessment. Similarly, there was some preliminary evidence that the sensitivity difference between men and women mainly reflects a sensitivity difference in judgments about given decisions rather than in standard assessment. This claim, if further corroborated, would also question the idea that gender differences in sensitivity as observed in attributions to bias (Stereotypical Bias Condition) and in attributions to error in committee decisions (Placement Condition) reflect selection bias; it is difficult to see how selection bias could differentially affect these conditions and the Decision Condition. Conversely, these results suggest that the effect of committee decision on response bias found in the BDT reflects an effect on standard assessment and might therefore be completely unrelated to actual committee decisions. In other words, this response-bias effect might not reflect participants' differential reactions to favoring versus disfavoring, or more versus less harmful, observed acts of discrimination; instead, it might reflect an effect of the standard (i.e., a tendency to perceive employees as generally deserving a pay raise rather than not) against which decisions are held.

General Discussion

In the present work, we investigated attributions to gender discrimination using a new paradigm, the Bias Detection Task (BDT), which allows researchers to apply Signal

Detection Theory (SDT) to such judgments and thereby differentiate between the ability to accurately detect discrimination (i.e., sensitivity) and the inclination to make more liberal or conservative judgments about the presence or absence of discrimination (i.e., response bias). In the gender BDT, participants are presented with a number of pay raise decisions involving male and female employees and have to judge whether they indicate signs of bias. We collected data totaling $N = 1704$ participants in different conditions with the goal to characterize male and female participants' response bias and sensitivity in attributions to gender discrimination. Our research questions were focused on the effects of cultural stereotypes and base rates (RQ 1), effects of harm and differences between favoring and disfavoring discrimination (RQ 2), and gender differences related to male and female participants' overall response bias and sensitivity (RQ 3) as well as intergroup effects on response bias (RQ 4).

Our results provide compelling evidence that participants' response bias reflects both cultural stereotypes about gender bias as well as actual base rates of different types of discrimination against men and women in specific contexts. In addition, we showed that this pattern reflects cultural stereotypes *about* and base rates *of gender bias* rather than participants' *own* gender biases. Sensitivity, on the other hand, was unrelated to such cultural stereotypes and base rates. In addition, the harm of a decision affected participants' response bias such that they were more liberal to judge more harmful decisions as being biased. This was the case both for their judgments about given decisions as well as for their judgments about who does versus does not deserve a pay raise delivered in the absence of committee decisions, questioning current accounts relating the effect to differences in judgments about favoring versus disfavoring or more versus less harmful acts of discrimination. We also found evidence for a corresponding sensitivity effect, such that sensitivity was larger for harmful, disfavoring discrimination, although that effect was considerably smaller than the response bias effect. Regarding sensitivity, our results showed higher sensitivity of men than women, contradicting our theoretical arguments for the

opposite pattern. In addition, we found no support for an alternative explanation of the observed opposite effect in terms of stereotype threat. Moreover, we did not find any evidence for ingroup-favoritism in participants' response bias. Our data did, however, provide some support for the idea that female participants have a more liberal response bias than male participants.

Substantive Implications

RQ 1: Base Rates and Cultural Stereotypes

Our results are in line with previous findings on the effects of base rates (Barrett & Swim, 1998; Inman, 2001) on attributions to discrimination but go beyond those findings in a number of ways. Most importantly, the BDT paradigm allows us to characterize base rate effects in terms of response bias – in line with hypotheses by Barrett and Swim (1998) about people's expectations about base rates and SDT applications in other contexts (Kellen & Klauer, 2018). Furthermore, in the BDT participants learn about base rates of discrimination *indirectly*, whereas previous studies often instructed participants directly about them (Inman, 2001; Major, Quinton, et al., 2002). Our results show that participants (a) learn about base rates online (as indicated by their ratings), (b) adapt their response bias accordingly, and (c) that this contingency learning depends on the context of discrimination (i.e., this pattern was absent in the Placement Condition which was not associated with discrimination). In addition to effects of discrimination base rates, response bias shifted in line with cultural stereotypes about gender discrimination as well. This was the case in the absence of gender discrimination in committee decisions and even when the pattern of discrimination in the committee decisions was contrary to stereotypes. However, sensitivity effects did not align with cultural stereotypes. Presumably, both the effect of base rates and cultural stereotypes on response bias are mediated by participants' expectations – a priori expectations through cultural stereotypes and expectations acquired online through base rates – which in turn lead to a shift in their response bias in the direction of their

expectations (Barrett & Swim, 1998; Inman, 2001; Kellen & Klauer, 2018).

RQ 2: Harm

The present results also align with the documented effects of harm on attributions to discrimination (Simon et al., 2019; Swim et al., 2003; York, 1989). Moreover, our work again extends those previous findings by characterizing such effects as predominantly reflecting differences in response bias with relatively small differences in sensitivity (cf. Barrett & Swim, 1998). Importantly, the absent difference in response bias effect between judgments about bias and about deservingness (see Experiment 4) suggests that the effect does not reflect differences in participants' judgments about observed acts of favoring versus disfavoring discrimination (Phillips & Jun, 2022). Instead, these effects can be explained by an overall bias to perceive potential victims as deserving positive treatment. However, we cannot completely discard the idea of a discrimination favorability effect on response bias based on our results. In our study, we explicitly instructed participants that discrimination entailed both unfairly denied and unfairly granted pay raises, which probably raised their expectations to encounter favoring discrimination, possibly diminishing according differences in response bias. Meanwhile, the corresponding sensitivity effect cannot be explained by a similar mechanism related to participants' own judgmental biases and therefore could reflect both observed harm as well as differences in judgments about observed favoring and disfavoring discrimination. To disentangle these factors, future studies could orthogonally manipulate the outcome of a decision (harmful vs. beneficial) and the framing (disfavoring vs. favoring) orthogonally in future studies.

RQ 3 and RQ 4: Gender Differences in Sensitivity and Response Bias

Our results revealed a less conclusive pattern regarding differences in attributions to discrimination between male and female observers. Contrary to common theoretical considerations, our studies consistently revealed higher sensitivity among male versus female participants. This difference was neither affected by the direction of gender discrimination

(Experiment 2) nor by the general context of discrimination (Experiment 3), which contradicts possible explanations of this effect in terms of a decrease in sensitivity for female participants through stereotype threat (related to either the experimental context threatening female employees or the general discrimination context in which women are more threatened than men). This left us with selection bias as a viable explanation, such that a difference in our participant pool of men and women causes the sensitivity effect. However, selection bias is difficult to reconcile with the fact that the sensitivity effect appeared only among participants' judgments in attribution tasks (i.e., the BDT and the Placement Condition), and not in judgment tasks (i.e., the Decision Condition). Still, more research is needed to confirm and explain such a difference.

Regarding response bias, our data provide some evidence that female observers have a more liberal response bias than male observers in the context of gender discrimination. This finding aligns with our arguments that women have higher perceived base rates of discrimination in general because of more frequent exposure (Kellen & Klauer, 2018), or increased vigilance to cues for discrimination (Higgins, 1996). Meanwhile, there was little evidence for ingroup-favoring patterns on response bias, as would be predicted by motivated attributions to discrimination for negative outcomes of an ingroup and positive outcomes of an outgroup member (Kaiser et al., 2006).

In general, the question remains as to why we did not find more robust gender differences in accordance with previous findings (Kaiser et al., 2006; Major et al., 2003; Major, Quinton, et al., 2002), given that there is consensus that men and women differ profoundly in their experiences with gender discrimination which in turn shapes their judgments in this context (Major, Quinton, et al., 2002). One possible explanation is related to individual differences: Although members of marginalized and privileged groups differ in their exposure to discrimination on average, there are considerable individual differences in exposure to discrimination within those groups (e.g., Becker & Swim, 2011), rendering group membership only a coarse proxy for experiences with discrimination. Furthermore, there is

evidence for the importance of several other predictors of individual differences in attributions to discrimination such as justice-related beliefs (e.g., Major, Gramzow, et al., 2002) and group identification (e.g., Major et al., 2003).

Methodological Implications and Outlook

Beyond the substantive implications elaborated above, the present work has methodological implications. With the BDT, we have introduced a paradigm that enables researchers to estimate sensitivity and response bias in attributions to discrimination which can readily be applied to a variety of other research questions. We demonstrated how separating sensitivity and response bias can generate new insights in the domain of attributions to gender discrimination. By manipulating employee characteristics, and characteristics of the committee and context, the BDT can provide further insights regarding the role of response bias and sensitivity for a variety of research questions.

As already elaborated, manipulating employee characteristics through the presentation of different portraits and names allows researchers to study attributions to discrimination based on different social categories. This positions researchers, for instance, to extend our results on gender discrimination by studying *racial discrimination*: Racial and gender discrimination differ in their origins and expression, and in corresponding differences in people's attributions, concerning, for instance, the effects of harm and intent (Simon et al., 2019). Thus, applying the BDT in the context of race might be worthwhile.

Furthermore, actor characteristics can be easily manipulated in the BDT. Whereas in our studies the composition of the decision committee was unknown, participants could be explicitly instructed about characteristics of the committee members. Doing so would, for instance, allow researchers to investigate the role of response bias and sensitivity in people's judgments as a function of discrimination prototypes (Inman & Baron, 1996; O'Brien et al., 2008; Simon et al., 2013). When situational features match people's expectations about typical actor, victim, and situation characteristics in instances of discrimination, they make

more attributions to discrimination. One such feature is status asymmetry: the expectation that actors from a privileged group discriminate against victims from a marginalized group. Thus, manipulating group status of committee members and employees accordingly in the BDT could give insights as to what extent this effect reflects differences in people's sensitivity or response bias.

Manipulating the context of the BDT would also allow researchers to study another feature of the discrimination prototype, stereotype asymmetry: the expectation that victims are negatively discriminated against in contexts where their group is negatively stereotyped (O'Brien et al., 2008). For instance, setting the paradigm in the context of a software development company (where women are negatively stereotyped) versus a childcare facility (where men are negatively stereotyped) would permit researchers to study the role of response bias and sensitivity in the context of stereotype asymmetry related to gender discrimination.

Finally, the BDT can be used to extend findings regarding individual differences in attributions to discrimination (Major, Quinton, et al., 2002). As already elaborated, our hypotheses regarding differences between members of marginalized and privileged groups are largely based on the assumption that members of different groups differ in their exposure to discrimination. Although this is certainly the case on average, exposure to discrimination varies considerably *within* those groups, and accounting for such variability may lead to more fine-grained predictions on the individual level (Becker & Swim, 2011; Major, Quinton, et al., 2002). Including measures of individual exposure to prejudice and discrimination in models of participants' response bias and sensitivity in the BDT would provide more informative tests of these hypotheses. Taken together, we hope to have not only contributed to the analysis of attributions to gender discrimination through our substantive results, but also by introducing a flexible and broadly applicable new paradigm for the investigation of attributions to discrimination and their analysis in terms of sensitivity and response bias.

Constraints on Generality

We sampled White native English-speakers from Western countries as participants in all experiments and presented portraits of White men and women to avoid potential interactions between gender discrimination and racial discrimination. Thus, the generalizability of our results to gender discrimination in populations of different ethnicities, to racial discrimination and to the interplay of both is an empirical question which we aim to examine in future studies. Furthermore, the present studies were based on the simplified binary conceptualization of gender. The extent to which our results generalize to attributions to discrimination involving nonbinary and gender-non-conforming individuals remains an open question because gender discrimination and discrimination based on queerness (independent of gender) are likely relevant. Finally, one of our central arguments (i.e., RQ 1 about the impact of cultural stereotypes) is based on Western cultural stereotypes about the nature and direction of gender discrimination in the context of corporate pay raise decisions. Consequently, we do not necessarily expect our results to generalize to non-Western populations where such stereotypes differ or to different workplace contexts (in particular considering the impact of context typicality on attributions to discrimination, as elaborated above). We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

References

Adams, G., Tormala, T. T., & O'Brien, L. T. (2006). The effect of self-affirmation on perception of racism. *Journal of Experimental Social Psychology*, *42*(5), 616–626. https://doi.org/10.1016/j.jesp.2005.11.001

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Association, A. P. (2022). Discrimination: What it is and how to cope. In *https://www.apa.org*. https://www.apa.org/topics/racism-bias-discrimination/types-stress

Aust, F., & Barth, M. (2023). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://github.com/crsh/papaja

Axt, J. R., Nguyen, H., & Nosek, B. A. (2018). The Judgment Bias Task: A flexible method for assessing individual differences in social judgment biases. *Journal of Experimental Social Psychology*, *76*, 337–355. https://doi.org/10.1016/j.jesp.2018.02.011

Barreto, M., & Ellemers, N. (2005). The burden of benevolent sexism: How it contributes to the maintenance of gender inequalities. *European Journal of Social Psychology*, *35*(5), 633–642. https://doi.org/10.1002/ejsp.270

Barreto, M., & Ellemers, N. (2015). Detecting and Experiencing Prejudice. In *Advances in Experimental Social Psychology* (Vol. 52, pp. 139–219). Elsevier. https://doi.org/10.1016/bs.aesp.2015.02.001

Barrett, L. F., & Swim, J. K. (1998). Appraisals of Prejudice and Discrimination. In *Prejudice* (pp. 11–36). Elsevier. https://doi.org/10.1016/B978-012679130-3/50036-3

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Becker, J. C., & Swim, J. K. (2011). Seeing the Unseen: Attention to Daily Encounters With Sexism as Way to Reduce Sexist Beliefs. *Psychology of Women Quarterly*, *35*(2), 227–242. https://doi.org/10.1177/0361684310397509

Blau, F., & Kahn, L. (2016). *The Gender Wage Gap: Extent, Trends, and Explanations. w21913*, w21913. https://doi.org/10.3386/w21913

Crocker, J., & Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review*, *96*(4), 608–630. https://doi.org/10.1037/0033-295X.96.4.608

Crosby, F., Clayton, S., Alksnis, O., & Hemker, K. (1986). Cognitive biases in the perception of discrimination: The importance of format. *Sex Roles*, *14*(11-12), 637–646. https://doi.org/10.1007/BF00287694

Crosby, J. R. (2015). The Silent Majority: Understanding and Increasing Majority Group Responses to Discrimination. *Social and Personality Psychology Compass*, *9*(10), 539–550. https://doi.org/10.1111/spc3.12196

Fitts, P. M. (1966). Cognitive aspects of information processing: III. Set for speed versus accuracy. *Journal of Experimental Psychology*, *71*(6), 849–857. https://doi.org/10.1037/h0023232

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*(3), 491–512. https://doi.org/10.1037/0022-3514.70.3.491

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Higgins, E. T. (1996). *Knowledge Activation: Accessibility, Applicability, and Salience*.

Inman, M. L. (2001). Do You See What I See?: Similarities And Differences In Victims' And Observers' Perceptions Of Discrimination. *Social Cognition*, *19*(5), 521–546. https://doi.org/10.1521/soco.19.5.521.19912

Inman, M. L., & Baron, R. S. (1996). Influence of prototypes on perceptions of prejudice. *Journal of Personality and Social Psychology*, *70*(4), 727–739. https://doi.org/10.1037/0022-3514.70.4.727

Jones, K. P., Peddie, C. I., Gilrane, V. L., King, E. B., & Gray, A. L. (2016). Not So Subtle: A Meta-Analytic Investigation of the Correlates of Subtle and Overt Discrimination.

Journal of Management, *42*(6), 1588–1613. https://doi.org/10.1177/0149206313506466

Kaiser, C. R., Vick, S. B., & Major, B. (2006). Prejudice Expectations Moderate Preconscious Attention to Cues That Are Threatening to Social Identity. *Psychological Science*, *17*(4), 332–338. https://doi.org/10.1111/j.1467-9280.2006.01707.x

Kellen, D., & Klauer, K. C. (2018). Elementary Signal Detection and Threshold Theory. In J. T. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (1st ed., pp. 1–39). Wiley. https://doi.org/10.1002/9781119170174.epcn505

König, R., & Heine, A. (2023). Learning to detect sexism: An evaluation of the effects of a brief video-based intervention using ROC analysis. *Frontiers in Psychology*, *13*, 1005633. https://doi.org/10.3389/fpsyg.2022.1005633

Lenth, R. V. (2023). *Emmeans: Estimated marginal means, aka least-squares means.* https://CRAN.R-project.org/package=emmeans

Major, B., Gramzow, R. H., McCoy, S. K., Levin, S., Schmader, T., & Sidanius, J. (2002). Perceiving personal discrimination: The role of group status and legitimizing ideology. *Journal of Personality and Social Psychology*, *82*(3), 269–282. https://doi.org/10.1037/0022-3514.82.3.269

Major, B., Quinton, W. J., & McCoy, S. K. (2002). Antecedents and consequences of attributions to discrimination: Theoretical and empirical advances. In *Advances in Experimental Social Psychology* (Vol. 34, pp. 251–330). Elsevier. https://doi.org/10.1016/S0065-2601(02)80007-7

Major, B., Quinton, W. J., & Schmader, T. (2003). Attributions to discrimination and self-esteem: Impact of group identification and situational ambiguity. *Journal of Experimental Social Psychology*, *39*(3), 220–231. https://doi.org/10.1016/S0022-1031(02)00547-4

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314–1334. https://doi.org/10.1037/a0012702

O'Brien, L. T., Kinias, Z., & Major, B. (2008). How status and stereotypes impact attributions to discrimination: The stereotype-asymmetry hypothesis. *Journal of Experimental Social Psychology*, *44*(2), 405–412. https://doi.org/10.1016/j.jesp.2006.12.003

Phillips, L. T., & Jun, S. (2022). Why benefiting from discrimination is less recognized as discrimination. *Journal of Personality and Social Psychology*, *122*(5), 825–852. https://doi.org/10.1037/pspi0000298

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rodin, M. J., Price, J. M., Bryson, J. B., & Sanchez, F. J. (1990). Asymmetry in prejudice attribution. *Journal of Experimental Social Psychology*, *26*(6), 481–504. https://doi.org/10.1016/0022-1031(90)90052-N

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. https://doi.org/10.3758/BF03196750

Salvatore, J., & Shelton, J. N. (2007). Cognitive Costs of Exposure to Racial Prejudice. *Psychological Science*, *18*(9), 810–815. https://doi.org/10.1111/j.1467-9280.2007.01984.x

Schmitt, M. T., Branscombe, N. R., Postmes, T., & Garcia, A. (2014). The consequences of perceived discrimination for psychological well-being: A meta-analytic review. *Psychological Bulletin*, *140*(4), 921–948. https://doi.org/10.1037/a0035754

Simon, S., Kinias, Z., O'Brien, L. T., Major, B., & Bivolaru, E. (2013). Prototypes of Discrimination: How Status Asymmetry and Stereotype Asymmetry Affect Judgments of Racial Discrimination. *Basic and Applied Social Psychology*, *35*(6), 525–533. https://doi.org/10.1080/01973533.2013.823620

Simon, S., Moss, A. J., & O'Brien, L. T. (2019). Pick your perspective: Racial group membership and judgments of intent, harm, and discrimination. *Group Processes & Intergroup Relations*, *22*(2), 215–232. https://doi.org/10.1177/1368430217735576

Singmann, H., & Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. In D. Spieler & E. Schumacher (Eds.), *New Methods in Cognitive Psychology* (1st ed., pp. 4–31). Routledge. https://doi.org/10.4324/9780429318405-2

Stangor, C., Swim, J. K., Van Allen, K. L., & Sechrist, G. B. (2002). Reporting discrimination in public and private contexts. *Journal of Personality and Social Psychology*, *82*(1), 69–74. https://doi.org/10.1037/0022-3514.82.1.69

Stroebe, K., Ellemers, N., Barreto, M., & Mummendey, A. (2009). For better or for worse: The congruence of personal and group outcomes on targets' responses to discrimination. *European Journal of Social Psychology*, *39*(4), 576–591. https://doi.org/10.1002/ejsp.557

Swim, J. K., Scott, E. D., Sechrist, G. B., Campbell, B., & Stangor, C. (2003). The role of intent and harm in judgments of prejudice and discrimination. *Journal of Personality and Social Psychology*, *84*(5), 944–959. https://doi.org/10.1037/0022-3514.84.5.944

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

York, K. M. (1989). Defining sexual harassment in workplaces: A policy-capturing approach. *Academy of Management Journal*, *32*(4), 830–850. https://doi.org/10.2307/256570