On Bias in Detecting Bias: A Signal Detection Analysis of Attributions to Gender Discrimination

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

[1]Albert-Ludwigs-Universität Freiburg

[2]University of California at Riverside

Accepted for Publication in the

Journal of Experimental Psychology: General

Word count: 15973

Author Note

The research in this article was supported by Grant Kl 614/40-1 from the German Research Foundation awarded to Karl Christoph Klauer. Marie Jakob received support from the research training group Statistical Modeling in Psychology (Grant 310365261), which is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG). The materials, data, and analysis scripts can be found at https://osf.io/6cgxv/?view_only=370994d8777d48cd8989a488f7f27c29.

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

Discrimination is the unjust treatment of people based on their group membership, and detecting discrimination is a necessary first step in countering it. Attributions to discrimination have often been conceptualized in terms of Signal Detection Theory (SDT), with researchers speculating about the role of sensitivity (i.e., observers' ability to recognize discrimination) and response bias (i.e., their tendency to judge that discrimination is present) in observers' judgments. Here, we take a first step toward moving beyond conceptual applications to empirical SDT analyses of attributions to discrimination across five studies examining judgments about gender discrimination in corporate pay raise decisions. We find that prototypes (Studies 1 to 3) and their alignment with the actual pattern of decisions (Studies 2 to 4) inform observers' response bias, reflecting differences in their attribution of unfair outcomes to discrimination (Study 5). In contrast, the extent of harm caused by discriminatory outcomes exerts a small effect on sensitivity and a large effect on response bias (Studies 1 to 3), which reflects observers' determination of a fair standard for decisions (Study 5). Theoretically, the present work contributes to the causal analysis of central factors shaping attributions to discrimination. Specifically, we shed light on whether effects of these factors are mediated by response bias or sensitivity, and assess their causal origin in either the application of a merit standard or attributional processing.

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

Public Significance Statement

This research investigates the perception of gender discrimination in corporate pay raise decisions. We replicate findings that more discrimination is perceived when decisions match prototypical expectations about gender discrimination and when decisions inflict harm. Using new methods we distinguish between different causes for these effects: (a) differences in people's ability to detect discrimination versus their tendency to assume it has occurred (causal pathway) and (b) differences in people's determination of a fair standard versus their judgment of an unfair decision as discriminatory (causal origin). Our studies reveal that stereotypes and harm influence the perception of discrimination through different mechanisms, which has theoretical and practical implications for understanding and improving the detection of discrimination.

Discrimination is the unfair or prejudicial treatment of people and groups based on characteristics such as race, gender, age or sexual orientation. Detecting discrimination is a necessary first step in countering it. However, the evidence for or against discrimination in specific cases is often ambiguous (e.g., Barreto and Ellemers, 2015; Crocker and Major, 1989; F. Crosby et al., 1986; Jones et al., 2016; Salvatore and Shelton, 2007; Schmader et al., 2022). According to Major, Quinton et al. (2002), "an attribution to discrimination is a judgment with two components: (1) the individual (or group) was treated unjustly, and (2) the treatment was based on social identity/group membership" (p. 262). Standards for what constitutes fair or just treatment are, however, often not available. Discrimination can also be expressed as unjustified differential treatment of people based on their group membership (Major, Quinton et al., 2002), but the contrast cases required to recognize differential treatment are often not available.[1] Additionally, attributing an outcome to discrimination requires an attribution of its cause to group membership. However, group membership is often confounded with various other variables.[2] Finally, many contemporary liberal societies place a strong emphasis on combating prejudice and discrimination. Consequently, when discrimination occurs, it is often hidden and disguised.

Given this background, the attribution of an outcome to discrimination frequently occurs in a context of considerable uncertainty. Modeling decisions under uncertainty is the goal of signal detection theory (SDT), and in a classical paper, Barrett and Swim (1998) used SDT as a theoretical lens to analyze attributions to discrimination. Originating from psychophysics, SDT was developed to describe how judgments about the presence of a physical signal (e.g., an acoustic tone) are made, but has since been used to model the detection of various types of psychological signals (Kellen & Klauer, 2018). A central contribution of SDT is its ability to separate two central components of decision-making: sensitivity, the ability to distinguish between

[1] For example, we may witness an immigrant being treated unfriendly by a civil servant, but may not know how the civil servant treats non-immigrants.

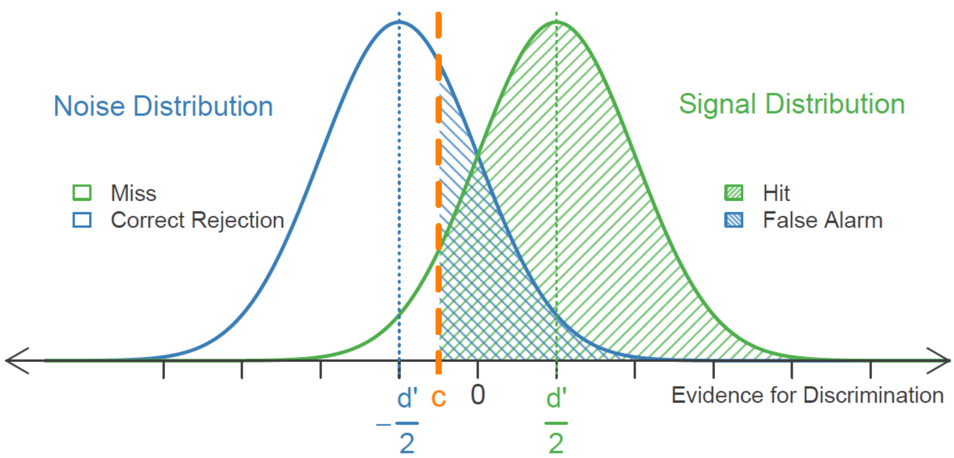
[2] For example, a Muslim person may not be selected as the new tenant of an apartment because of their religion or because the successful non-Muslim candidate was able to document more solid personal finances.

signal and noise, and response bias, a tendency to prefer one type of judgment over the other. In the context of attributions to discrimination, sensitivity refers to the ability to distinguish between cases with and without evidence of discrimination; It reflects the ability to correctly perceive and assess the relative strength of evidence of discrimination. Response bias, on the other hand, refers to the tendency to attribute an outcome to discrimination rather than to other possible causes.

Figure 1

Illustration of the SDT Model and the Types of Attributions

Attribution: Is Discrimination Present?	*True State:* Is Discrimination Present?	
	No (Noise)	Yes (Signal)
No („Noise“)	*Correct Rejection*	*Miss*
Yes („Signal“)	*False Alarm*	*Hit*



Note. Signal and noise distributions are assumed to be normal distributions with equal variance (set equal to one without loss of generality).

As a conceptual framework, SDT distinguishes four different types of attribution: false alarms, correct rejections, hits, and misses, depending on the true presence of a discrimination (i.e., the signal in this context) and the decision maker's judgment about it (see Figure 1). SDT assumes that the probability of these judgments depends on latent evidence of discrimination relative to a decision criterion *c*. An attribution to discrimination, that is, a "signal present" judgment, is made only if the evidence of discrimination surpasses this threshold. Evidence for signal (i.e., discrimination present) and noise (discrimination absent) events are modeled through two distributions separated by d', which provides a measure of sensitivity. As d' increases, both false alarms and misses become less likely. The response criterion c constitutes a measure of response bias. As c increases, the probability of misses increases while the probability of false alarms decreases, implying more conservative response bias (and vice versa as c decreases). In other words, a larger, more conservative response bias implies that a decision maker needs more evidence to make an attribution to discrimination (relative to a smaller, more liberal response bias).

Empirical SDT Analyses of Attributions to Discrimination

SDT has been a useful theoretical lens for research on attributions to discrimination because it is well understood how central elements of decisions under uncertainty affect response bias or sensitivity, and thereby the pattern of the different types of attribution shown in Figure 1. Central elements include the costs of different types of decision, perceived base rates, signal strength (the strength of evidence of discrimination), and the availability and mobilization of attentional and cognitive resources. Research on attributions to discrimination has accumulated a wealth of knowledge on these elements, often directly referring to SDT: on costs incurred in attributions to discrimination (Adams et al., 2006; Barreto and Ellemers, 2015; Branscombe et al., 1999; Major et al., 2007; Pinel, 2004); on perceived base rates as shaped by perceived prevalence and shared prototypes (Barreto and Ellemers, 2015; Inman and Baron, 1996; Inman et al., 1998; Miller and Saucier, 2018; O'Brien et al., 2008; Simon et al., 2013); on the relative evidential strength of different cues to discrimination (Simon et al., 2019; Swim et al., 2003; York, 1989); and on the deployment of attentional and cognitive resources (Allport, 1954; Kaiser et al., 2006).

For these reasons, the seminal paper by Barrett and Swim (1998) on SDT was highly influential, and researchers continue to use SDT as a framework to theorize about attributions to discrimination to this day. For example, Heineck and Deutsch

(2025) recently argued that when victims of disadvantaging actions are perceived as prototypical, this “likely introduces a response bias in discrimination judgments. Following the Signal Detection Theory, a tendency to preferentially acknowledge false alarms while striving to avoid misses should be expected when observing prototypicality is high” (p. 3).

However, to our knowledge, no study that has gone beyond using SDT as a conceptual framework for theorizing about the mediation[3] of effects through response bias or sensitivity to conducting *empirical* SDT analyses of attributions to discrimination (Merritt, 2026).[4] The present research takes a first step towards filling this gap.

Empirical SDT analyses contribute to theory by clarifying mediation pathways. To illustrate, consider the well-established finding that members of lower-status groups are more likely to attribute negative outcomes experienced by members of their group to discrimination than members of higher-status groups are (e.g., women more than men, Swim et al., 2003; Black people more than White people, Simon et al., 2019; Latin-American people more than White people; Adams et al., 2006; see also J. R. Crosby, 2015). This effect could reflect differences in sensitivity or differences in response bias. Two factors may boost the sensitivity of members of lower-status groups: In line with an idea going back to Allport’s (1954) vigilance perspective, they may invest more attentional resources to monitor events for signs of discrimination (Kaiser et al., 2006). In addition, frequent exposure to acts of discrimination may have made members of lower-status groups experts in detecting evidence of discrimination. Increased monitoring for signs of discrimination and learning through frequent

[3] Here, we use the term "mediate" in a conceptual rather than a statistical sense. That is, we use it to refer to the idea that an effect may operate through one of several possible causal pathways (e.g., via an effect on response bias) rather than through another one (e.g., via an effect on sensitivity) within a theoretical framework.

[4] Note, however, that König and Heine (2023) performed what they called an SDT analysis in the context of judging statements sexist versus non-sexist. They compute an index of sensitivity known as the area under the curve (Kellen & Klauer, 2018), which they incorrectly interpret as a composite measure comprising the influence of sensitivity and response bias. However, they do not disentangle response bias from sensitivity, which is arguably the key advantage of SDT analyses in the context of attributions to discrimination.

exposures may thus increase sensitivity of lower-status group members. However, a case can also be made for response bias: Frequent exposure to acts of discrimination may raise the perceived base rate of discrimination for members of lower-status groups, and base rates are a factor known to inform response biases in many domains (Barrett & Swim, 1998; Kellen & Klauer, 2018). People from lower-status groups may also sometimes be motivated to attribute negative outcomes to discrimination in order to to deflect blame from their ingroup onto the moral shortcomings of the higher-status group (Crocker and Major, 1989, but see Major, Gramzow, et al., 2002; Major et al., 2003), leading to lowered response thresholds. Finally, heightened vigilance might make discrimination a highly accessible construct (Higgins, 1996) raising not only the likelihood that evidence of discrimination is interpreted as such, but also the likelihood that innocuous or ambivalent features of situations are construed as cues to discrimination.[5] Vigilance, motivated attributions, and perceived base rate might therefore lead to a more liberal[6] response bias for members of lower-status groups.

An empirical SDT analysis could unravel these theoretical possibilities. Such an analysis would clarify whether the effect is a sensitivity effect, mediated by an effect on the observers' ability to perceive discrimination, or a response bias effect, mediated by an effect on the observers' inclination to make attributions to discrimination.

A New Paradigm: The Bias Detection Task

Empirical SDT analyses require a study design in which many scenarios with evidence of discrimination (signal trials) and scenarios without evidence of discrimination (noise trials) are presented. However, most existing studies have relied on presenting a few scenarios, often only one (e.g., Barreto and Ellemers, 2015; Inman, 2001; Inman and Baron, 1996; O'Brien et al., 2008; Phillips and Jun, 2022; Simon et al., 2019; Stroebe et al., 2009); noise trials without evidence of discrimination are only rarely included (but see Inman and Baron, 1996; König and Heine, 2023; Swim et al.,

[5] Overall additive shifts of perceived evidence strength of this kind are subsumed under response bias effects in SDT.

[6] By "liberal" ("conservative") we mean response bias favoring (discouraging) an attribution to discrimination.

2003). Therefore, we devised a new paradigm, the Bias Detection Task, that administers a large number of cases with (i.e., signal) and without (i.e., noise) evidence of discrimination and thereby enables empirical SDT analyses.

In most of the studies we report here, the Bias Detection Task is set in the context of pay raise decisions:[7] Participants are told that employees in the marketing department of a large company are evaluated for possible pay raises. Decisions are made by a panel of supervisors based on scores that characterize an employee's performance on three job-relevant criteria (see Figure 3). The supervisor panel typically bases its decisions on the scores, but it has additional information at its disposal (employee CVs, employment history, etc.), which may in special cases weigh in on the decision. This creates a typical ambiguity regarding the causes of specific outcomes and affords the supervisor panel discretionary latitude, which may give rise to discriminatory decisions (Doering et al., 2023).

The merit standard that determines whether an employee deserves a pay raise is defined through a decision rule based on the employee's criterion scores. Participants learn the merit standard and how to apply it in the learning phase at the beginning of the Bias Detection Task. They are then told that there have been complaints that the panel of supervisors has made a number of biased decisions and that biased decisions can lead to certain employees being denied a pay raise although they deserve one, and certain employees being granted a pay raise although they do not deserve one. In response to the complaints, the company has decided to organize an independent review of the decisions. Participants are told that they are part of an initial screening of decisions based only on the most basic information of each case (i.e., the scores). Their task "is to flag those decisions that you suspect might be biased and that should receive further scrutiny."

In line with these instructions, noise trials (i.e., trials without cues to discrimination) represent cases in which an employee is denied a pay raise and does not deserve one, or an employee is granted a pay raise that they deserve. These decisions

[7] The rationale of the Bias Detection Task can, however, readily be applied to other contexts as exemplified in the studies reported below.

are objectively just and by definition free of discrimination. Signal trials with cues to possible discrimination are those in which an employee is denied a pay raise although they deserve one or granted a pay raise although they do not deserve one.

For each case, participants are presented with portraits and names of each employee, the employee's scores and the decision of the panel. Hence, cues to group membership (e.g., in terms of gender or race) can be added (see lower panel of Figure 3). In the present studies, we manipulate the gender of the employees, keeping their ethnicity constant.[8]

Attributions to Gender Discrimination in the Bias Detection Task

To establish the Bias Detection Task we focus on effects of harm, victim prototypicality and observer gender, as these are among the most influential variables in previous research on attributions to discrimination (Barrett & Swim, 1998; Inman & Baron, 1996; Swim et al., 2003). Even more importantly, these variables have been frequently discussed in relation to SDT (e.g., Barrett & Swim, 1998; Heineck & Deutsch, 2025; Merritt, 2026), allowing us to generate hypotheses about how they may affect sensitivity and response bias in attributions to discrimination. Our expectations for how these factors and their effects would manifest in the SDT analysis were additionally informed by a pilot study with the Bias Detection Task (described in the introduction to Study 1).

Effects of Harm: Sensitivity or Response Bias?

Firstly, prior findings suggest that decision makers consider the harm of a discriminatory behavior (Simon et al., 2019; Skarlicki & Kulik, 2004; Swim et al., 2003; York, 1989) and make more attributions to discrimination when stronger harm is incurred. Thus, in the Bias Detection Task, we would expect participants to make more attributions to discrimination in cases in which a pay raise was denied (which causes considerable harm to the employee in question) than in cases in which a pay raise was

[8] The present work focuses on gender discrimination in the context of a binary perception of gender, without thereby endorsing a binary conceptualization of gender more broadly. Rather, our work builds on previous research on gender discrimination focused on men and women, and may be extended in future work to include individuals with gender identities beyond these binary categories.

granted (which does not harm but rather benefits the employee). Effects of harm may be mediated through response bias or sensitivity: On the one hand, Barrett and Swim (1998) argued that the intensity of cues to discrimination increases as more harm is incurred, and that increases in intensity decrease ambiguity, allowing for greater sensitivity. However, Swim et al. (2003) asserted that harmful consequences of behaviors and decisions increase the likelihood of attributions to discrimination even in the absence of clear cues to discrimination such as an intent to discriminate. This is consistent with the idea that the effects of harm on judgments of discrimination are mediated by response bias. Similarly, Barrett and Swim's (1998) arguments on how differential costs for false alarms and misses shape response bias also suggest a response bias effect given that a miss (a biased decision is overlooked and accepted) is more costly than a false alarm (an unbiased decision is questioned as biased) for pay-raise denied decisions and vice versa for pay-raise granted decisions. Additionally, cases in which a pay raise was denied versus granted may reflect instances of disfavoring (i.e., unfairly receiving a more negative treatment than deserved because of one's group membership) versus favoring (i.e., unfairly receiving a more positive treatment than deserved because of one's group membership) discrimination, respectively. Recent evidence suggests that participants make fewer attributions to discrimination when presented with cases displaying favoring versus disfavoring discrimination (Phillips & Jun, 2022). This similarly suggests that people attribute more to discrimination when a pay raise is denied than when it is granted and could also be mediated through response bias or sensitivity.

Victim Gender and Prototypicality

Beyond harm, victim prototypicality is known to affect attributions to discrimination, with prototypical victims being more likely to elicit such attributions (Carlsson & Sinclair, 2018; Heineck & Deutsch, 2025; Inman & Baron, 1996; O'Brien et al., 2008; Rodin et al., 1990). In the context of the present studies (i.e., gender discrimination in corporate pay raise decisions) shared cultural knowledge of, for example, the gender pay gap, suggests that women constitute prototypical victims of disfavoring discrimination (taking the form of unjust "denied" decisions in the Bias Detection Task), whereas men may be considered prototypical beneficiaries of favoritism (taking the form of "granted" decisions; Carlsson & Sinclair, 2018; Harris et al., 2004). Effects related to victim prototypicality may again be mediated through sensitivity or response bias: Heineck and Deutsch (2025) assert that prototypical victims may shift decision maker's response bias. On the other hand, Inman (2001) suggested that expectations related to prototypicality are sensitivity factors.

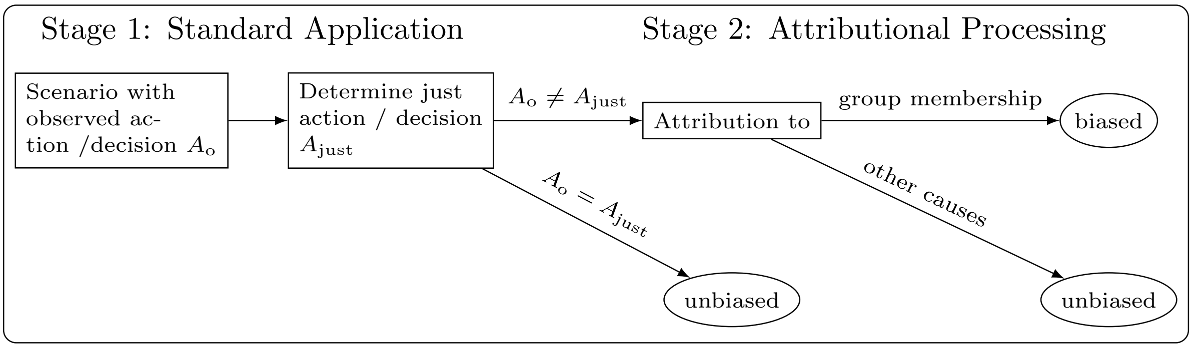
Observer Gender Differences

Finally, as elaborated above, male and female observers may differ in their sensitivity or response bias due to their different experiences with gender discrimination. More precisely, heightened vigilance for gender discrimination, 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. However, our pilot study surprisingly revealed an opposite tendency: Male participants' sensitivity was larger than that of female participants. In the following studies, we test a number of different explanations for this result, focusing on characteristics of our participants and their decision processes, selection bias, as well as differential threat experienced by male and female participants, which could differentially impair their performance (Schmader et al., 2008).

Standard Application and Attributional Processing

Apart from clarifying mediation pathways of effects through sensitivity and response bias, a second theoretical contribution of this work is to disentangle the origin of effects according to the two components of attributions to discrimination described by Major, Quinton et al (2002). Recall that according to their definition, an attribution to discrimination entails (1) that an individual was treated unfairly and (2) that the cause of this treatment was their group membership. This translates to two decision stages, illustrated in Figure 2: *standard application* and *attributional processing.* In standard application, observers first determine and apply a merit standard in order to evaluate

Figure 2
Decision Stages in Attributions to Discrimination



Note. The figure depicts an adaptation of Major, Quinton et al.'s (2002) analysis of attributions to discrimination for the present paradigm. In our paradigm, and in many cases in real life, decision makers apply a merit standard to determine what the fair or just decision / action is. This self-determined decision, A_{just}, then serves as input to further attributional processing that can end with an attribution to discrimination depending upon the match ($A_o = A_{just}$) or mismatch ($A_o \neq A_{just}$) to the observed (panel) decision, A_o, and the perceived causes of possible mismatches.

what a just decision (A_{just}) would have been. In attributional processing, observers then determine whether this just decision matches the observed one (A_o) and, in the case of a mismatch ($A_o \neq A_{just}$), consider possible causes of the mismatch (such as the victim's group membership). To illustrate, consider a case in the Bias Detection Task in which a female employee with above-average scores on all criteria is denied a pay raise (i.e., the observed outcome A_o). According to the framework sketched above (see Figure 2), in the standard application stage an observer would first determine whether or not the employee deserves a pay raise and thus determine a (hypothetical) fair outcome A_{just} (e.g., because the employee has above-average scores on all criteria, the observer reckons the employee would have deserved a pay raise). Then, they would assess if the fair outcome they determined (A_{just}) matches the actual outcome (A_o, the denial of a pay raise in this case). In case of a match, they would judge the observed decision as unbiased (since no unfair treatment occurred in their opinion). If the hypothetical fair outcome and the actual outcome do not match, the observer would consider the given decision as unfair. They could either attribute the cause of this unfair treatment to the employee's group membership and judge the case as biased (i.e., make an attribution to discrimination) or attribute to other causes, resulting in an "unbiased" judgment.

Differentiating standard application and attributional processing is possible through a comparison of the Bias Detection Task with a related task, the Judgment Bias Task, which measures social biases and also utilizes SDT (Axt et al., 2018). Both tasks present a number of scores along with cues to social group membership (gender categories in the present case) and both require a decision based on the scores. The central difference between them is that in the Bias Detection Task, participants judge given decisions as biased or unbiased, whereas in the Judgment Bias Task, they make those decisions themselves (i.e., there are no panel decisions). Accordingly, the Judgment Bias Task aims to measure participants' *own biases* whereas the Bias Detection Task assesses differences in participants' *attributions to bias* for given decisions.

Participants' judgments in the Bias Detection Task, where their task is to judge decisions as biased, are determined by the outcomes of both standard application and attributional processing, (as illustrated above). Meanwhile, in the Judgment Bias Task, participants simply make what they consider an appropriate decision without making any subsequent attributions. In the above example, participants would simply use the information presented in the case (i.e. the employee's gender and their scores on the criteria) to decide whether or not the employee in question deserved a pay raise, without making any further attributions. Therefore, participants' decisions in the Judgment Bias Task reflect only standard application.

Thus, contrasting the Bias Detection Task with the Judgment Bias Task allows us to determine the extent to which effects in participants' attributions to discrimination in the Bias Detection Task originate during standard application or reflect differences in attributional processing. This comparison is the objective of Study 5.[9]

[9] Note that this decomposition according to the two stages of standard application and attributional processing, like any decomposition, rests on assumptions that we critically address in the General Discussion.

Table 1

Definitions of Key Terms.

Term	**Definition**
Signal Detection Theory (SDT)	A theoretical and statistical framework for studying human decisions under uncertainty. Used here to separate response bias and sensitivity in attributions to discrimination.
Response Bias	One of the two central SDT parameters, which captures a tendency to prefer one type of judgment (in this case, an attribution to discrimination) over the other.
Sensitivity	One of the two central SDT parameters, which describes the ability to distinguish between signal (i.e., the presence of cues to discrimination) and noise (i.e., the absence of cues to discrimination).
Bias Detection Task	A novel paradigm to study attributions to discrimination that, in conjunction with Signal Detection Theory, can be used to separate the contributions of response bias and sensitivity to those attributions.
Judgment Bias Task	A paradigm developed by Axt et al. (2018) to measure individuals' social biases.
Standard Application	The first stage of an attribution to discrimination according to the definition by Major, Quinton et al. (2002): Observers first apply a standard to determine a just action or decision.
Attributional Processing	The second stage of an attribution to discrimination according to the definition by Major, Quinton et al. (2002): Observers determine whether the action they determined to be just matches the observed action. In case of a mismatch, consider possible causes for the mismatch, such as the victim's group membership.

Theoretical and Practical Significance

The present research aims to take a first step towards empirical SDT analyses of attributions to discrimination with two overarching goals: (1) clarifying whether effects are mediated through sensitivity or response bias, and (2) identifying their causal origin in the standard application or attributional processing stage. Both distinctions carry theoretical and practical significance. Theoretically, response bias and sensitivity reflect distinct cognitive processes. A sensitivity effect implies that a given variable changes observers' ability to perceive and evaluate evidence of discrimination, whereas a response bias effect implies that the variable shifts the threshold beyond which the same evidence is interpreted as discriminatory. Identifying which mechanism underlies a given effect therefore constrains theoretical accounts of attributions to discrimination in ways that conventional paradigms cannot. This decomposition is grounded in SDT and has been discussed since the seminal paper by Barrett and Swim (1998). However, to date it has never moved beyond conceptual applications of SDT. The distinction between standard application and attributional processing similarly distinguishes cognitive processes: an effect originating in standard application implies that a variable shapes observers' standard for fair behavior, whereas an origin in attributional processing implies that it affects how observers interpret the cause of an unfair decision. This decomposition thereby introduces a novel theoretical dimension to causal analyses in this field.

Practically, both distinctions have implications for the design of interventions to improve discrimination detection. Training programs aimed at sensitizing observers to detect bias (Becker & Swim, 2011) would be appropriate when effects are driven by sensitivity, as they aim to enhance the ability to perceive diagnostic cues. However, when effects are driven by response bias, interventions would need to address the threshold at which evidence is acted upon, for instance, by correcting beliefs about the prevalence of discrimination. Similarly, effects driven by standard application call for interventions that improve how merit is evaluated, whereas effects driven by attributional processing require interventions targeting causal reasoning about unfair outcomes.

The Present Research

As a first empirical step towards these goals, we present five studies with our novel paradigm, the Bias Detection Task, focusing on effects of harm, victim prototypicality, and observer gender in attributions to gender discrimination. Study 1 establishes an empirical foundation of these effects for the subsequent studies. On this basis, Study 2 manipulates the direction of the panel's gender bias to investigate the relative contribution of victim prototypicality and actual discrimination patterns on observers' response bias. Study 3 further explores sensitivity effects through a high-powered design. Study 4 then implements a condition without any cues to discrimination to examine how far the response bias effects of harm and victim prototypicality, as well as the observer gender difference in sensitivity, depend on culturally shared knowledge about discrimination. Across Studies 2 - 4, we also test explanations for the sensitivity difference between male and female participants in terms of different types of threat. Finally, Study 5 investigates whether effects of the factors manifest in standard application versus attributional processing.

Study 1

As outlined above, the goal of Study 1 is to confirm pilot results regarding effects of harm, victim prototypicality, and observer gender on attributions to discrimination, providing an empirical foundation for the subsequent studies. Study 1 and the pilot study both presented cases involving male and female employees, sampled male and female participants, and panel decisions tended to favor men and disfavor women (see the Online Appendix for the exact results of the pilot study).

Findings of the Pilot Study

The results of the pilot study served as preregistered hypotheses for Study 1. Regarding response bias, the results revealed that (1) participants' response bias in the pilot study differed as a function of the panel decision: They were more likely to consider decisions biased when a pay raise was denied rather than granted, corresponding to a more liberal response bias for harmful than beneficial panel decisions.

This finding is in line with the documented effects of harm (see above, e.g., Phillips and Jun, 2022; Simon et al., 2019; Skarlicki and Kulik, 2004; Swim et al., 2003; York, 1989). The main effect was moderated by (2) an interaction with employee gender: Participants were more inclined to judge panel decisions biased if they disfavored female rather than male employees and favored male rather than female employees. This finding is in line with prior findings on victim prototypicality (Heineck & Deutsch, 2025; Inman & Baron, 1996). Additionally, (3) the two-way interaction was moderated by participant gender such that this pattern was partially reversed for male participants. This pattern corresponds to a relative ingroup bias because each group tended to consider harmful decisions more likely biased if they involve ingroup rather than outgroup targets, and vice versa for beneficial decisions. Being response bias effects, the above effects affect cases with and without evidence of discrimination equally.

Regarding sensitivity, theoretical notions such as vigilance and expertise through exposure reviewed above would lead one to expect higher sensitivity and/or a more liberal response bias for women than for men. Interestingly, as already mentioned above, (4) men tended to be more sensitive than women.

These findings of the pilot study were the preregistered hypotheses for Study 1.

Methods

Transparency and Openness

All studies reported in this article were pre-registered on the OSF, where we also provide all data and analysis scripts for this project (see Table 1). Studies 4 and 5 were preregistered as one study with two conditions, but we report them as two studies for the sake of clarity. Note that in doing so we did not deviate from any of the preregistered hypotheses or analyses, since our preregistered plan involved only contrasts between each condition and the stereotypical bias condition, and not between each other.

We obtained an ethics approval from our university's ethics committee for all studies within this project before any data were collected. We report all manipulations, measures, and exclusions in these studies. Deviations from the preregistered plans are noted in the main text; the Online Appendix includes a table detailing all deviations.

Sample Size Justification and Power

In all studies, the target sample size was 120 men and 120 women per between-participant condition of the respective study design. As preregistered, we collected complete data from 125 men and 125 women per condition to account for exclusions and included all eligible datasets in our analyses. The target sample size was determined by a power analysis based on effect-size estimates from the pilot study.

The effects found in the pilot study had effect sizes of, in order, $d_z = 0.17$, $d_z = 0.65$, and $d_z = 0.93$ for the effects on response bias; the effect of participant gender on sensitivity had an effect size of $d = 0.33$. The target sample size of 240 participants (120 men; 120 women) per group yields very high power (larger than .95) to detect effects with $d_z = 0.65$ and $d_z = 0.93$ and reasonable power for the smaller effect sizes, .75 (.84) for $d_z = 0.17$ and .72 (.81) for the between-participant effect with $d = 0.33$, in two-tailed (one-tailed) tests of undirected (directed) hypotheses.

Table 1

Open Materials, Open Data, and Preregistrations

Study	Content	URL
All	Materials, Data, Analysis Scripts, Online Appendix	https://osf.io/6cgxv/?view_only=370994d8777d48cd8989a488f7f27c29
1	Preregistration	https://osf.io/f473b/?view_only=8bc88f99353148d6be04c705a942504d
2	Preregistration	https://osf.io/qwfe4/?view_only=cd2766065ae44dd39e903c74a13f0705
3	Preregistration	https://osf.io/su9xn/?view_only=39605967d5bf4dc5a911cf73aaf84bfc
4 & 5	Preregistration	https://osf.io/e7fw4/?view_only=c2f961f70800439988b2f934d7fc62f5

Note. Although Studies 4 and 5 were preregistered as one study with two conditions, we report them as two studies for the sake of clarity. Note that in doing so we did not deviate from any of the preregistered hypotheses or analyses, since we contrasted each condition with the stereotypical bias condition and not with each other.

Participants

The experiment was conducted as an online experiment on Prolific, as were the other studies reported in this manuscript. In all our studies, participants had to be White[10] 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 and had to have already participated in at least five studies on Prolific. As preregistered, 250 participants took part. Three participants were excluded (two because their mean sensitivity was close to chance level,[11] indicating that they did not work properly on the task, and one because they reported an ethnicity other than White), resulting in a final sample size of N = 247 (122 women). The mean age of the sample was $M = 34.03$ years ($SD = 7.51$), ranging from 19 to 49 years.

Procedure

The experiments took participants on average 50 minutes to complete. They first completed the Bias Detection Task, for which they received detailed instructions; important parts of the instructions were tested by three comprehension checks. The participants had two chances to read the relevant instructions and correctly respond to the questions for each comprehension check. If they did not respond correctly on their second attempt, the experiment immediately ended for them. The Bias Detection Task comprised a learning phase and an evaluation phase.

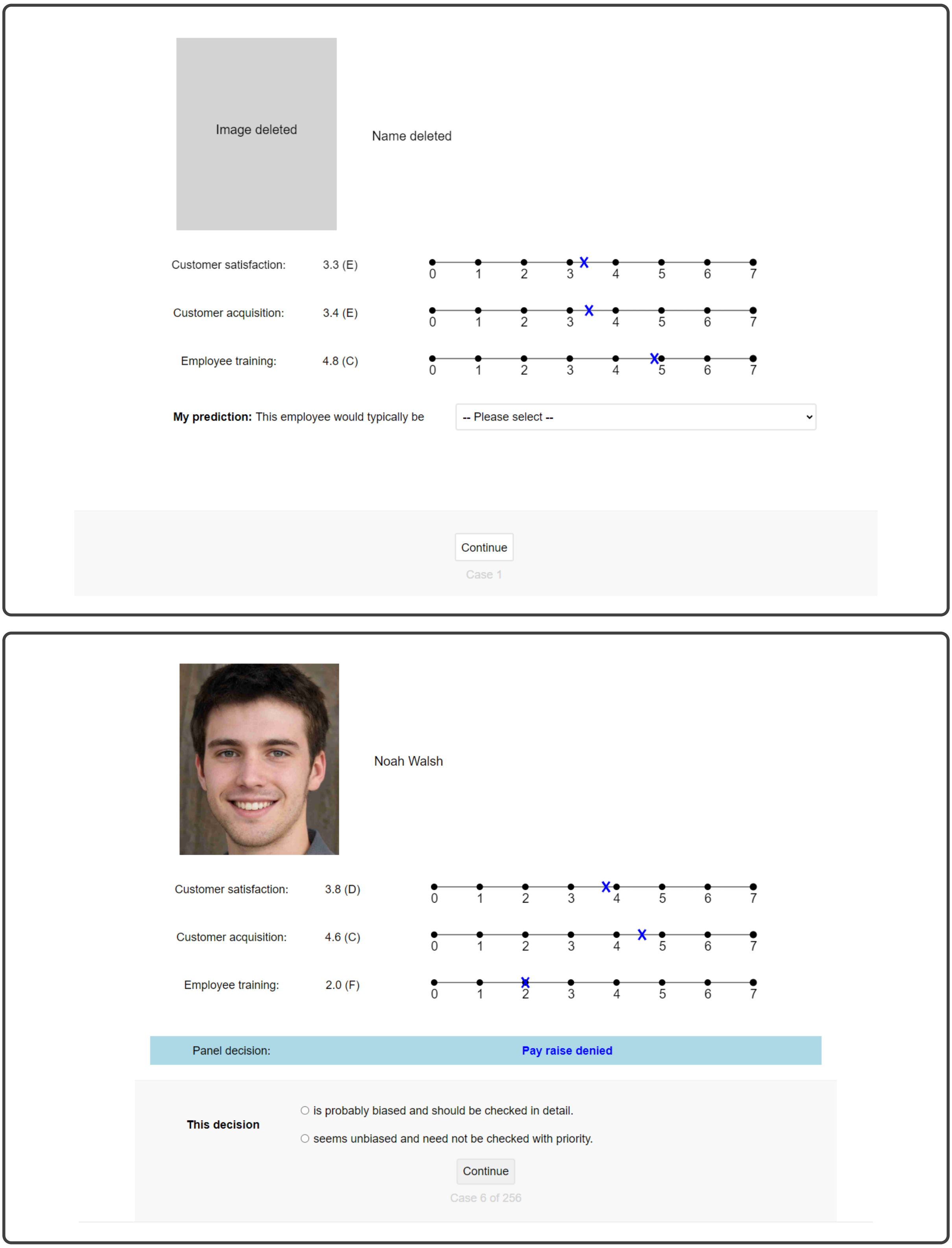
Learning Phase. In the learning phase, participants' task is to familiarize themselves with how the scores typically translate into panel decisions. The cases are presented in an anonymized format (see upper panel of Figure 3). For each case, participants have to predict the panel decision based on the employee's scores. They are then shown the actual panel decision, which always respected the decision rule in this phase. Unknown to participants, the decision rule was that an employee deserves and is

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

[11] Although we exclude participants with overall sensitivity (in terms of d') smaller than 0.1 in all of our studies on suspicion of not having engaged with the task and always intended to do so, we forgot to preregister this exclusion criterion.

Figure 3

Example Trials of the Learning (Upper Panel) and Evaluation Phase (Lower Panel)



Note. In the learning phase, participants predict the panel decision based on the employee's scores and are then shown the panel decision that always respects the merit standard that is to be practiced in the learning phase. In the evaluation phase, participants see portraits and names of each employee along with the scores and the panel decision. The cues (portrait and name) to group membership in terms of gender are redacted in learning trials.

granted a pay raise if the mean of their three scores is above 3.5 (the scale midpoint), and does not deserve and is denied one if it falls below. The purpose of the learning phase is to familiarize participants with this decision rule, which then serves as their standard for judging decisions as just or unjust. The learning phase ended once participants has predicted 18 out of 20 consecutive decisions correctly.

Evaluation Phase. After the learning phase, participants were told that there had been complaints about biased pay raise decisions and that in response, the company had implemented a review of the decisions. Participants were asked to help in this review by identifying cases that are likely biased for further in-depth scrutiny. Due to the large number of decisions to be reviewed, they were told that they would see only cases that were relatively close calls and would see only basic information (i.e., scores and basic demographic information) for each case.

In the evaluation phase, cues to group membership are added through the presentation of a portrait picture and the name of the employee (see the lower panel of Figure 3). Participants respond to 256 trials, indicating in each case whether the panel decision is likely biased and needs further scrutiny or not.

After the evaluation phase, participants rated aspects of their perception of the panel decisions. Specifically, they were asked about the frequency with which male and female employees were unfairly granted and denied a pay raise, respectively, the panel's gender bias and the extent to which it discriminated against women and men (see the Online Appendix and Tables 3 and 4 for details). There were no preregistered analyses for the ratings and descriptive results are reported in the Online Appendix. Finally, participants had the opportunity to describe their strategy in the Bias Detection Task, answered a number of demographic questions, and were asked if they had responded seriously and attentively so that their data could be used for analysis.

Design

Each trial in the evaluation phase specified the panel decision, the employee gender, and, via the scores, whether the employee did or did not deserve a pay raise according to the merit standard practiced in the learning phase. Consequently, within-participant factors were "panel decision", "employee gender", and "merit". As shown in Table 2, the panel decisions exhibited gender bias in that they sometimes unjustly disfavored women and favored men.

Table 2

Stimulus Design: Frequencies of Trial Types as a Function of Merit (Pay Raise Deserved vs. Not Deserved), Panel Decision (Granted vs. Denied), and Employee Gender (Female vs. Male) in the Evaluation Phase

	Pay Raise Deserved		Not Deserved	
Employee Gender	Granted	Denied	Granted	Denied
— Studies 1 and 2: Stereotypical Gender Bias —				
Female	32	32[a]	0[a]	64
Male	64	0[a]	32[a]	32
— Study 2: Reversed Gender Bias —				
Female	64	0[a]	32[a]	32
Male	32	32[a]	0[a]	64
— Study 3: No Gender Bias —				
Female	40	24[a]	24[a]	40
Male	40	24[a]	24[a]	40

Note. Panel decisions in Study 1 exhibited gender bias in that half of the women deserving a pay raise were denied one, whereas half of the men not deserving a pay raise were granted one. In Study 2, this pattern was reversed in the group with reversed bias so that women were favored and men disfavored. In Study 3, the number of unjust decisions of either kind (pay raise denied although deserved, and pay raise granted although not deserved) was the same, but there was no correlation with gender. Note that the within-participants factors are not fully crossed in Study 1 and 2 as evidenced by the zero cells.
[a]Unjust decisions.

Materials

In all experiments, we used AI-generated portrait pictures of 128 White male and 128 White female faces (Karras et al., 2019). Portraits were selected from a larger pool of portraits in a validation study based on ratings of the age and attractiveness of the depicted persons and on how typical they are for a White person, a Black person, a man and a woman. We selected portraits that were (a) unambiguously classified in terms of ethnicity and gender and such that (b) men and women were perceived as of similar age and attractiveness. Details can be found in the Online Appendix.

The scores of each employee on the criteria (see Figure 2) were drawn from a multivariate normal distribution with a mean of 3.5 and moderate correlations between the criteria. For the evaluation phase, scores were sampled so that each employee's mean fell within the middle 50 % of the implied score distribution — that is, between the first and third quartile. This was consistent with the instruction that the screening was restricted to close calls (mean scores were close to the 3.5 cut-off value) and prevented the task from being too easy. For the learning phase, scores were not constrained.

Data Analysis

SDT provides a measure of response bias in terms of the response criterion c and a measure of sensitivity d' (see Figure 1). We estimated these parameters using a mixed-effects maximum likelihood approach. With this approach, effects of independent variables and their interactions on sensitivity and response bias can be modeled as fixed effects, while differences between participants in their overall sensitivity and response bias as well as in the effect of the fixed effects are modeled through random effects (see Jakob et al., 2025, for an introduction to mixed-effects SDT and Singmann & Kellen, 2019, for a general introduction to mixed-effects modeling). For significance testing, we relied on model comparisons with χ^2 likelihood ratio tests (for confirmatory analyses) and Wald tests reported as z statistics (for exploratory analyses).

Our model included fixed effects on response bias for the factors panel decision, employee gender, participant gender, and all of their interactions. Regarding sensitivity, our design (see above) allowed us to estimate either the fixed effect of panel decision or employee gender.[12] We report the results of a model including a fixed effect on

[12] As described above, our design is not fully crossed (i.e., signal cases never involved a female employee being granted, nor a male employee being denied, a pay raise). Therefore, it was not possible in this design to estimate sensitivity effects of employee gender and panel decision simultaneously, nor

sensitivity for employee gender; the alternative model including the panel decision instead produced similar results and the same substantive conclusions. The method for determining the random-effects structure was preregistered, the emerging models for each experiment are described in the Online Appendix.

Absolute Effect Sizes. Parameter values for response bias and sensitivity are represented on the shared strength-of-evidence dimension (see Figure 1). They are standardized because the variance of the signal distribution and the noise distribution are set to 1 (without loss of generality), and SDT provides a transparent translation between the parameter values and the response rates of the different types of attribution. Therefore, we specify effect sizes directly as differences on the strength-of-evidence dimension, referred to as Δ in the following.

Results

The fixed intercept for sensitivity was $\beta_0^{d'} = 1.66$ ($z = 38.9$, $p < .001$) showing that participants were generally able to distinguish between signal and noise. Regarding response bias, the fixed intercept of $\beta_0^c = 0.15$ ($z = 8.99$, $p < .001$) indicates a tendency towards an "unbiased" relative to a "biased" judgment.

Figure 4 shows the response bias parameters as a function of panel decision and employee gender. Effect sizes and their 95% confidence intervals are shown in Figure 7.

Preregistered Analyses

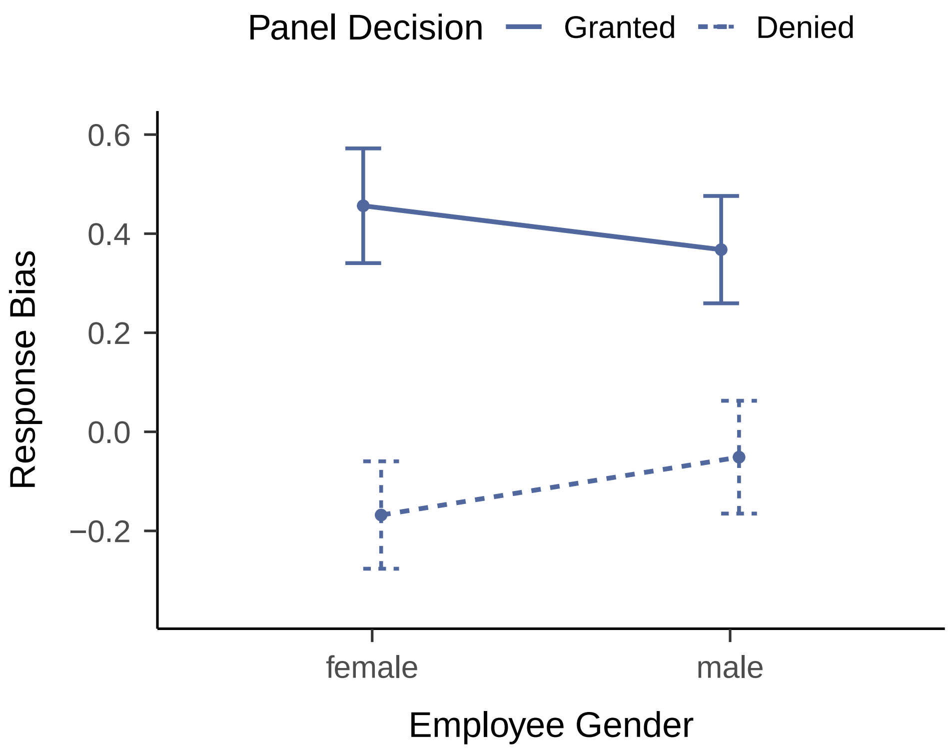
Response Bias. Considering the preregistered hypotheses, there was (1) an effect of panel decision on response bias. Participants were more likely to consider harmful (pay raise denied) decisions biased than beneficial (pay raise granted) ones ($\chi^2(1) = 23.34$, $p < .001$, $\Delta = 0.522$, 95% CI $[0.315, 0.728]$).

(2) The effect was moderated by employee gender. Participants were more inclined to judge pay-raise denied decisions biased for female than for male employees and vice versa for pay-raise granted decisions ($\chi^2(1) = 27.10$, $p < .001$, $\Delta = 0.103, 95\%$, CI $[0.065, 0.141]$). However, (3) this two-way interaction was not significantly moderated by participant gender in the direction of an ingroup bias ($\chi^2(1) = 2.88$, $p = .090$,

their interaction. Study 3 below eliminated this limitation.

Figure 4

Response Bias as a Function of Employee Gender and Panel Decision



Note. Large values of response bias correspond to conservative biases, that is, less of an inclination to consider a given panel decision "biased". Conversely, small values indicate liberal response biases. Error bars show 95% confidence intervals.

$\Delta = 0.033$, 95% CI $[-0.005, 0.071]$). Given the relatively small p value, we nonetheless followed up on this three-way interaction in Study 2 with higher statistical power.

Sensitivity. Regarding sensitivity, (4) female participants exhibited somewhat lower sensitivity than male participants ($\chi^2(1) = 6.63$, $p = .010$, $\Delta = 0.221$, 95% CI $[0.053, 0.388]$).

Exploratory Analyses

In addition to the preregistered analysis of the effect of the panel decision on response bias, we exploratorily tested for a corresponding effect on sensitivity, which was not significant, $z = 1.037$, $p = .300$, $\Delta = 0.0451$, 95% CI $[-0.0401, 0.130]$.[13] Furthermore, regarding the participant gender difference in sensitivity, we examined response times of male and female participants as a possible explanation, as speed-accuracy trade-offs might differ between male and female participants. Specifically, male participants may prioritize accuracy over speed to a greater extent than female participants, which would result in greater accuracy as well as longer response times. As members of a privileged group, men may be motivated to avoid being perceived as biased in a gender-based categorization task, which could lead to a stronger focus on accuracy. In contrast, female participants, as members of a marginalized group in terms of gender (which is particularly salient in the context of the task), may worry less about being perceived as biased and therefore prioritize accuracy less than men do.

However, contrary to this explanation, the correct responses of male participants tended to be *faster* than the ones of female participants ($M_{\text{male}} = 4.79$ seconds, $M_{\text{female}} = 5.03$ seconds, $F(1, 245) = 3.58$, $p = .059$, $\eta_p^2 = .059$).[14]

Discussion

The confirmatory tests of Study 1 found that panel decision, employee gender, and participant gender exerted effects on attributions to discrimination, reflecting differences in response bias in some cases, and differences in sensitivity in other cases.

Response Bias

The presented panel decision affected participants' response bias, which is well aligned with the documented effects of harm (Phillips & Jun, 2022; Simon et al., 2019; Skarlicki & Kulik, 2004; Swim et al., 2003; York, 1989). However, our results go beyond those previous findings by characterizing this effect as a difference in response bias: Harmful decisions were more likely to be judged potentially biased than beneficial decisions whether or not the employee was treated unjustly. At the same time, harm

[13] Note that different from the confirmatory analyses, our model for this analysis included a fixed effect of committee decision and no effect of employee gender on sensitivity.

[14] 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.

did not affect sensitivity, that is, participants' ability to discriminate between cases with and without evidence of discrimination. The present results thus tentatively resolve the theoretical possibilities outlined in the introduction in favor of effects of harm on response bias rather than sensitivity (Barrett & Swim, 1998; Swim et al., 2003).

The main effect of panel decision on response bias was moderated by an effect of employee gender in a two-way interaction. The participants' inclination to respond "biased" was increased if a female employee was denied a pay raise and a male employee was granted one relative to when a male employee was denied one and a female employee was granted one. On the one hand, this result is in line with effects of victim prototypicality on attributions to discrimination (Heineck & Deutsch, 2025; Inman & Baron, 1996), as the cases with a more liberal response bias are consistent with prototypes as commonly discussed, for example, in the context of the gender pay gap. Alternatively, however, participants may have taken into account the contingencies of the stimulus design (see Table 2), which was consistent with the prototype: female but not male employees were sometimes unjustly denied a raise, while male but not female, employees were sometimes unjustly granted one. In consequence, the pattern of response bias might also reflect a learned *contingency-based* response bias acquired online. We pursue this question further in the subsequent studies.

Sensitivity

One limitation of Study 1 is that its unbalanced design does not allow us to test whether the interaction of panel decision and employee gender affects participants' sensitivity as well. As already explained, prototypical expectations may not only affect response bias but sensitivity as well (Inman, 2001). We designed Study 3 to close this gap.

Considering participant gender, male participants exhibited somewhat greater sensitivity in distinguishing between cases with and without evidence of discrimination than female participants. This finding is surprising given that the theoretical accounts reviewed in the introduction would (if anything) predict greater sensitivity for women rather than men.

Our exploratory analysis of participants' response times already revealed evidence against a plausible explanation of that effect in terms of speed-accuracy trade-offs. In the following studies, we consider additional possible explanations of this finding in terms of threat and selection bias: Stereotypical gender bias in panel decisions disfavors women and therefore threatens women more than men. Threat has often been argued to have an impact on participants' performance (e.g. Schmader et al., 2008). Differential threat might therefore explain differences in sensitivity. We will test this *threat hypothesis* in the subsequent studies by implementing conditions that should reverse or remove the differential threat in different ways.

Another possible explanation is selection bias, which is a pervasive issue when groups cannot be formed by randomly assigning participants to them, as is obviously the case for gender groups. Such groups are sometimes referred to as non-equivalent (Shadish et al., 2002) to remind us of the fact that there is no guarantee that such groups differ only in the feature that defines the intended, focused-on difference (i.e., gender in the present case). For example, the reasons men and women choose to participate in Prolific studies may differ in ways that are largely unknown. Hence, it is likely that the gender groups differ in many characteristics other than gender (for example, we already reported evidence suggesting differences in response speed; the Prolific participant pool contains many more women than men; women aborted the experiment more often than men: 39.5 % vs. 23.8 % in Study 1; our male sample was older than the female sample: $M_{\text{female}} = 33.19$ years vs. $M_{\text{male}} = 34.86$ years).[15] Small overall differences in sensitivity may thus be due to the fact that the gender groups are not equivalent. Finding such sensitivity differences pervasively, and even in conditions that remove any hint of gender discrimination, would support the selection-bias hypothesis.

Taken together, Study 1 provides a first basis for effects of harm, victim prototypicality and observer gender on sensitivity and response bias in attributions to

[15] In the Online Appendix, we report results of exploratory analyses where we examined the extent to which the participant gender sensitivity difference is related to differences in demographic characteristics between the groups across all studies. None of the examined variables contributed to the effect.

discrimination. Key findings include that response bias reflected harm as well as victim prototypicality (through an interaction of harm and victim gender) and that sensitivity was greater for male than female observers. The following studies further explore different explanations and extensions of these effects.

Study 2

Study 2 explores potential mechanisms underlying two findings from Study 1: (1) the extent to which the interaction of harm and employee gender in response bias is prototype-based rather than contingency-based and (2) the influence of threat on the participant gender difference in sensitivity.

To this end, we contrasted a condition with the same stimulus design as in Study 1 (i.e., a replication of Study 1, referred to as *stereotypical bias condition*) with a new condition, the *reversed bias condition.* Here, the roles of male and female employees in the stimulus design were reversed, so that the panel now exhibited reversed or counter-stereotypical gender bias (see Table 2).

Regarding the interaction between harm and employee gender in response bias in Study 1, our design allows us to examine the possible contributions of contingency-based and prototype-based mechanisms in shaping this interaction. If it purely reflects contingency-based response bias, the interaction should occur with the same strength in the reversed bias as in the stereotypical condition, but in *reversed direction.* This pattern would indicate that in both conditions, participants learn the respective stimulus contingencies and adjust their response bias to reflect the correlation between panel decision and employee gender, which reverses from the stereotypical to the reversed bias condition (see Table 2). However, if the interaction is driven by a prototype-based mechanism, drawing on the shared cultural knowledge of how gender discrimination plays out in pay raises, then its strength and direction should be similar in both conditions.

Consider next the gender difference in sensitivity observed in Study 1. The reversed bias condition favors female employees and disfavors male employees and therefore, the implemented pattern of discrimination threatens women less than the stereotypical bias condition. In fact, male participants may even experience the reversed bias condition as relatively more threatening than female participants. A reduced or even reversed participant gender sensitivity difference in the reversed bias condition would therefore support the threat hypothesis.

Finally, because results regarding a possible three-way interaction of participant gender, panel decision, and employee gender were inconclusive, we also preregistered another test of the three-way interaction in the stereotypical bias condition as already foreshadowed above. This possible three-way interaction takes the shape of an ingroup bias in response bias.

Methods

Participants

As preregistered, the between-participants conditions of Studies 2, 4 and 5 (stereotypical bias, reversed bias, placement, judgment bias) were sampled in one data collection with random assignment of participants to the four conditions. 508 participants completed the stereotypical bias and the reversed bias condition. Of these, seven were excluded because their mean sensitivity was close to zero (see Footnote 11) and 14 because they reported being of a different ethnicity than White. Total sample size after exclusions was $N = 487$, with 245 participants in the condition with stereotypical bias (122 women), and 242 participants in the condition with reversed bias (120 women). The mean age of the sample was $M = 34.2$ years ($SD = 7.56$), ranging from 18 to 49 years.

Procedure

Study 2 followed the same procedure as Study 1 with the main difference that participants in the two conditions worked with different stimulus designs, one implying stereotypical gender bias, the other reversed bias, as specified in Table 2 (see Online Appendix for a description of a few additional minor procedural changes).

Results

The fixed intercepts were descriptively similar to Study 1 (Sensitivity: $\beta_0^{d'} = 1.78$, $z = 58.7$, $p < .001$; Response Bias: $\beta_0^{c} = 0.142$, $z = 10.56$, $p < .001$), again indicating above-chance discernment of signal and noise, as well as a tendency towards making "unbiased" judgments. Overall sensitivity ($z = 0.87$, $p = .386$) and response bias ($z = 0.85$, $p = .395$) did not differ significantly between conditions.

Figure 5 shows response bias as a function of panel decision, employee gender, and condition (stereotypical bias vs. reversed bias). Figure 7 shows effects sizes and their 95% confidence intervals for the within-group effects.

Replication

As preregistered, we first tested whether the effects of Study 1 were replicated in the condition with stereotypical bias. All findings were replicated:

- A strong main effect of panel decision on response bias, with harmful panel decisions more likely to be judged biased than beneficial ones: $\chi^2(1) = 9.74$, $p_{\text{one-tailed}} = .002$, $\Delta = 0.236$, 95% CI $[0.089, 0.382]$.
- A two-way interaction between panel decision and employee gender on response bias, with stereotype-congruent decisions (women-denied, men-granted) more likely to be judged biased than stereotype-incongruent ones (women-granted, men-denied): $\chi^2(1) = 41.58$, $p_{\text{one-tailed}} < .001$, $\Delta = 0.102$, 95% CI $[0.071, 0.132]$.
- The absence of a moderation of this interaction by participant gender: $\chi^2(1) = 0.08$, $p_{\text{one-tailed}} = .388$, $\Delta = 0.009$, 95% CI $[-0.053, 0.070]$.
- The main effect of participant gender on sensitivity, with female participants being somewhat less sensitive than male participants: $\chi^2(1) = 7.77$, $p_{\text{one-tailed}} = .005$, $\Delta = 0.250$, 95% CI $[0.075, 0.425]$.

Effects of Contingency Condition

Preregistered Tests. As can be seen in Figures 5 and 7, the two-way interaction of panel decision and employee gender found in the condition with stereotypical bias was reduced and close to zero, but not reversed in the condition with reversed bias. Statistically, this translated into (1) a significant two-way interaction between panel decision and employee gender ($\chi^2(1) = 20.92$, $p_{\text{one-tailed}} < .001$, $\Delta = 0.052$, 95% CI $[0.029, 0.073]$), which was (2) moderated by contingency condition in a significant three-way interaction ($\chi^2(1) = 19.88$, $p_{\text{one-tailed}} < .001$, $\Delta = 0.050$, 95% CI $[0.028, 0.072]$).

Figure 5

Panel Decision × Employee Gender Interaction in Response Bias in Experiment 2

Note. Large values of response bias correspond to conservative biases, that is, less of an inclination to consider a given panel decision “biased”. Conversely, small values indicate liberal response biases. Conditions of Study 2 are abbreviated as "Stereotypical" (stereotypical bias condition) and "Reversed" (reversed bias condition). Error bars show 95% confidence intervals.

Across both conditions, the sensitivity of male participants was again higher than that of female participants ($\chi^1(1) = 27.79$, $p_{\text{one-tailed}} < .001$, $\Delta = 0.323$, 95% CI

$[0.205, 0.442]$). This effect was not significantly moderated by contingency condition ($\chi^2(1) = 1.42$, $p_{\text{one-tailed}} = .884$, $\Delta = -0.073$, 95% CI $[-0.191, 0.046]$). Descriptively, the gender difference was more pronounced in the reversed bias condition than in the stereotypical bias condition (see Figure 7).

Exploratory Analyses.

Tables 3 and 4 show the means and standard deviations of the ratings participants provided at the end of the Bias Detection Task. Overall, the rating responses support the conclusion that participants picked up the direction of gender bias. All ratings differ significantly between the conditions with different directions of gender bias (smallest $t(484.8) = 11.11$, largest $p < .001$, smallest $d = 1.01$, 95% CI $[0.818, 1.20]$).

Discussion

The stereotypical bias condition was a replication of Study 1. The same pattern of effects emerged in this condition of Study 2 as in Study 1.

Response Bias

Moving from stereotypical bias to reversed bias, the two-way interaction between panel decision and employee gender was neither reversed, as a purely contingency-based response bias would predict, nor unchanged, as a purely prototype-based response bias would predict. This outcome suggests that both learning and culturally-shared prototypes shape response bias: On the one hand, the reduced two-way interaction and ratings of the panel's gender bias in the reversed bias group indicate that participants picked up the implemented contingencies, which shifted their response bias and the perception of the panel's bias. On the other hand, neither response bias nor ratings fully reflected the reversal of bias in the panel decisions, suggesting that prototypical expectancies contribute a component that learning did not fully overcome.

Sensitivity

Turning to sensitivity, we argued that female participants should have felt less threatened relative to male participants in the reversed than in the stereotypical bias condition. In fact, according to their ratings, male and female participants perceived

Table 3

Means and Standard Deviations (in parentheses) of Frequency Ratings in Studies 2, 3, and 4

Condition	Q1 (W Granted)	Q2 (W Denied)	Q3 (M Granted)	Q4 (M Denied)
Stereotypical Bias (Studies 2 and 4)[a]	2.20 (0.88)	3.54 (0.84)	3.58 (0.79)	2.31 (0.84)
Reversed Bias (Study 2)	3.08 (0.76)	2.71 (0.81)	2.68 (0.87)	3.14 (0.75)
No Bias (Study 3)	2.85 (0.73)	3.29 (0.72)	3.22 (0.71)	2.94 (0.69)
Placement (Study 4)	2.94 (0.68)	2.95 (0.63)	2.84 (0.68)	2.83 (0.64)

Note. Seven-point Likert scales asked how often women who did not deserve a pay raise were granted one (Q1, abbreviated as "W Granted") and how often women who deserved a pay raise were denied one (Q2, abbreviated as "W Denied"); Q3 (abbreviated as "M Granted") and Q4 (abbreviated as "M Denied") asked these questions for men. Response options ranged from 0 = 'Never' to 6 = 'Always'. In the placement condition, merit was replaced by relative aptitude and pay raise decision by placement decisions in these questions. See the online Appendix for the exact wordings.

[a]This condition was considered in Studies 2 and 4; see Methods sections.

less gender discrimination in the former condition than in the latter condition. Nevertheless, women were again somewhat less sensitive than men in discriminating between cases with and without evidence of discrimination and this gender difference was not reduced in the reversed bias condition. Thus, this initial test of the threat hypothesis does not provide support for it. Instead, the robustness of the gender difference is more consistent with an explanation based on selection bias.

Taken together, Study 2 thus establishes that the interaction of harm and employee gender in response bias reflects both prototype-based and contingency-based contributions. The threat hypothesis for the observer gender differences in sensitivity, however, did not receive support.

Table 4

Means and Standard Deviations (in parentheses) of Bias Ratings in Studies 2, 3, and 4

Condition	Q5 (Gender Bias)	Q6 (Discrimination)	Q7 (Favoritism)	Q8 (Harm)
Stereotypical Bias (Studies 2 and 4)[a]	3.01 (0.78)	2.99 (0.72)	3.04 (0.82)	2.96 (0.73)
Reversed Bias (Study 2)	1.62 (0.91)	1.64 (0.84)	1.55 (0.92)	1.73 (0.90)
No Bias (Study 3)	2.24 (0.81)	2.25 (0.77)	2.28 (0.81)	2.26 (0.73)
Placement (Study 4)	2.18 (0.60)	2.18 (0.60)	2.08 (0.60)	2.12 (0.52)

Note. Five-point Likert scales asked ratings of the direction and extent of the panel's gender bias (Q5), its discrimination (Q6), its favoritism (Q7), the harm it caused (Q8). The endpoint 0 indicates a response consistent with reversed gender bias; the midpoint 2 with the absence of systematic gender bias; endpoint 4 with stereotypical gender bias; See the Online Appendix for the exact wordings.

[a]This condition was considered in Studies 2 and 4; see Methods sections.

Study 3

Study 3 aimed to provide a more complete picture of possible sensitivity effects. To that end, it addresses a key limitation of Studies 1 and 2: Instead of an unbalanced stimulus design implementing stereotypical and reversed gender bias (as Studies 1 and 2 did, see Table 2), Study 3 implemented a balanced design with no systematic gender bias. This setup allows us to disentangle effects of employee gender, committee decision, and their interaction on sensitivity. Theoretical arguments lend a certain plausibility to such an interaction (Inman, 2001): For example, observers may scrutinize actions and decisions more carefully in cases where stereotypes suggest a greater likelihood of discrimination, which could lead to higher sensitivity for such cases. Alternatively, in line with the stereotype inconsistency effect in person memory research (Hastie & Kumar, 1979), observers may be especially attentive to cues for discrimination in prototype-inconsistent cases, which would suggest the opposite pattern in sensitivity.

With regard to response bias, results of Study 2 suggest that the interaction between panel decision and employee gender on response bias is shaped by both a stereotype-based and a contingency-based mechanism. In Study 3, the stimulus contingencies did not implement systematic gender bias and thus, the influence of the contingency-based component on this interaction was removed. Consequently, if an interaction emerges in Study 3, then it must reflect prototype-based response bias shaped only by a priori prototypes. This line of reasoning predicts an interaction of a size that is intermediate between the effects found in the stereotypical and reversed bias conditions in Study 2.

Moreover, the absence of gender bias in the panel's decision should again reduce perceptions of threat by female participants like in the reversed bias condition in Study 2. This permits another test of the threat hypothesis of gender differences in sensitivity. Because the threat hypothesis did not receive support in Study 2, we expected to again find the gender difference in sensitivity in Study 3.

Methods

Participants

In Study 3, the 256 trials are spread out across more cells of the stimulus design than in Studies 1 and 2, resulting in fewer data per cell. Overcompensating for the loss of power implied thereby, we doubled the target sample size, and aimed to recruit 240 men and 240 women. 511 participants completed the experiment. Of these, seven were excluded because their mean sensitivity was close to zero (see Footnote 11); 14 because they reported being of a different ethnicity than White; one because they were non-binary and two because they responded that their data should not be used. The total sample size after exclusions was $N = 487$ (246 women). Mean age was $M = 33.37$ years ($SD = 7.22$), ranging from 19 to 49 years.

Procedure

The procedure was the same as in Study 2, with the exception that a different stimulus design was used (see Table 2).

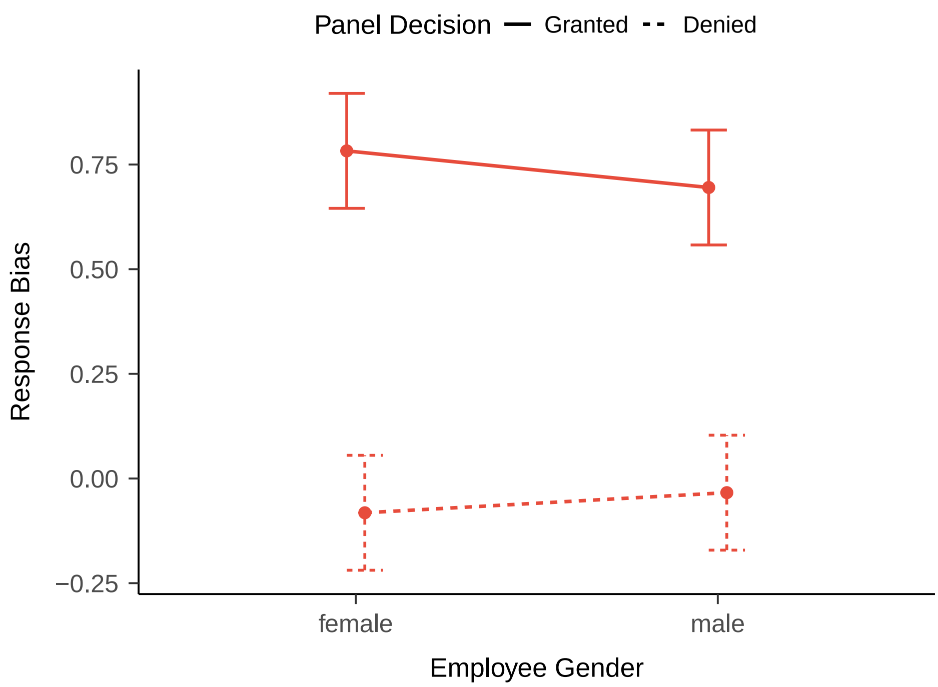
Results

Fixed intercepts for sensitivity and response bias were again similar to the previous studies (Sensitivity: $\beta_0^{d'} = 1.74$, $z = 63.2$, $p < .001$; Response Bias: $\beta_0^c = 0.17$, $z = 13.8$, $p < .001$).

Response Bias

Figure 6

Panel Decision × Employee Gender Interaction in Response Bias in Study 3 (No Bias)



Note. Large values of response bias correspond to conservative biases, that is, less of an inclination to consider a given panel decision "biased". Conversely, small values indicate liberal response biases. Error bars show 95% confidence intervals.

As hypothesized, there was again an effect of committee decision on response bias, which was additionally moderated by employee gender: Overall, participants' response bias was smaller for decisions where a pay raise was denied than when a pay raise was granted ($\chi^2(1) = 36.9$, $p_{\text{one-tailed}} < .001$, $\Delta = 0.406$, 95% CI $[0.278, 0.535]$). Additionally, as displayed in Figure 6, response bias was again more liberal in prototype-congruent cases (women being denied and men being granted a pay raise) than in incongruent cases ($\chi^2(1) = 12.13$, $p_{\text{one-tailed}} < .001$, $\Delta = 0.035$, 95% CI $[0.015, 0.054]$).

Moreover, an exploratory analysis revealed an effect of participant gender on response bias: Women's response bias was somewhat smaller (more liberal) than men's ($z = 3.08$, $p = .002$, $\Delta = 0.077$, 95% CI $[0.028, 0.126]$).

Sensitivity

Study 3 was designed to allow us to separately test main effects of panel decision, employee gender, and their interaction on sensitivity. Results revealed a main effect of panel decision on sensitivity, which was somewhat higher for harmful "denied" decisions than for beneficial "granted" decisions ($\chi^2(1) = 11.01$, $p < .001$, $\Delta = 0.072$, 95% CI $[0.030, 0.114]$). No significant effects emerged for he main effect of employee gender ($\chi^2(1) = 0.297$, $p = .586$, $\Delta = 0.010$, 95% CI $[-0.026, 0.046]$) or the interaction between panel decision and employee gender ($\chi^2(1) = 0.069$, $p = .793$, $\Delta = -0.005$, 95% CI $[-0.040, 0.031]$). Moreover, results again revealed the participant gender difference, with male participants exhibiting higher sensitivity than female participants ($\chi^2(1) = 6.08$, $p_{\text{one-tailed}} = .007$, $\Delta = 0.136$, 95% CI $[0.028, 0.243]$).

Ratings

As can be seen in Tables 3 and 4, participants' ratings of the panel's degree of gender bias were intermediate between the stereotypical bias and the reversed bias conditions of Study 2.

Discussion

Study 3 replicated key findings of Studies 1 and 2: There were again effects of panel decision and an interaction of panel decision and employee gender in response bias. Figure 7 displays effect sizes and confidence intervals for the central effects across all studies; it shows that the interaction in this study (implementing contingencies with no gender bias) was intermediate in size between the ones found in the two conditions of Study 2 (implementing stereotypical and reversed gender bias, respectively). This pattern of results is in line with the idea that response bias is shaped by both contingency- and prototype-based components, as the interaction in Study 3 reflects the prototype- but not the contingency-based component (which was removed through the balanced contingencies).

In addition, women were again somewhat less sensitive than men in distinguishing between cases with and without evidence of discrimination. This finding emerged despite the absence of systematic gender bias in the panel decisions, which should have reduced any threat experienced by women.

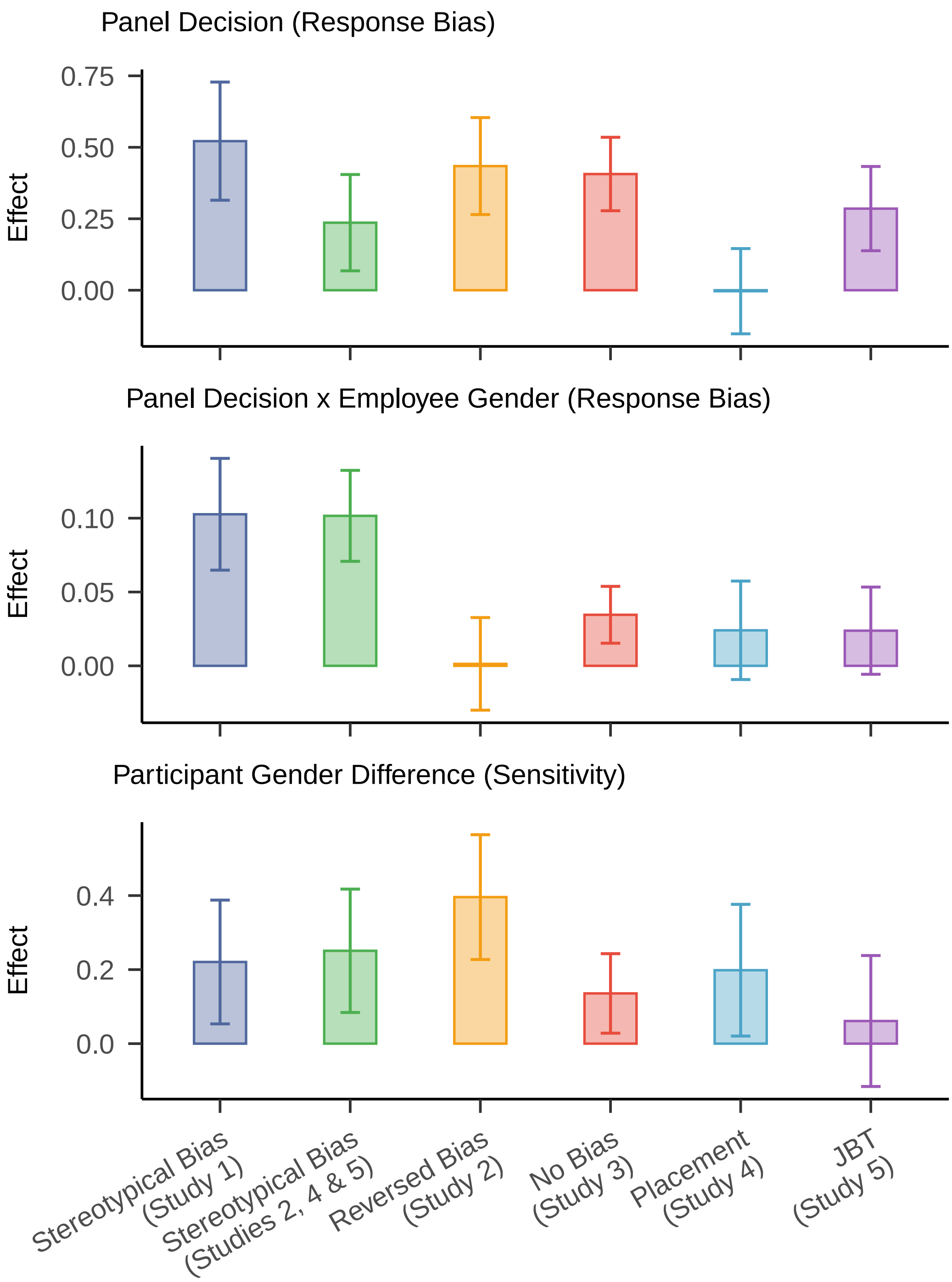
Study 3 also provided a more detailed and higher-powered analysis of the effects on sensitivity. This revealed an effect of panel decision on sensitivity such that participants' sensitivity was higher for cases with harmful than with beneficial panel decisions. In a joint analysis of Studies 1 to 3, the effect is also significant ($\chi^2(1) = 10.08$, $p = .0015$, $\Delta = 0.042$, 95% CI $[0.014, 0.069]$) and not significantly moderated by condition ($\chi^2(2) = 4.83$, $p = .090$). Taken together, harm appears to exert a robust and strong effect on response bias, and also affects sensitivity to a small degree, as originally surmised by Barrett and Swim (1998), possibly reflecting more thorough scrutiny of harmful decisions than of beneficial decisions.

Finally, an exploratory analysis revealed a more liberal response bias for women than men, which is in line with prior findings and the theoretical arguments reviewed in the Introduction. In a joint analysis of Studies 1 to 3, the effect is also significant ($z = 3.15$, $p = .002$, $\Delta = 0.053$, 95% CI $[0.021, 0.085]$) and not significantly moderated by condition ($\chi^2(2) = 1.69$, $p = .429$, $\Delta = 0.042$). Thus, our results appear to also provide evidence for an observer difference in response bias.

Taken together, Study 3 finds no evidence that victim prototypicality shapes sensitivity, suggesting that effects of victim prototypicality on attributions are mediated through response bias, rather than changes in sensitivity. Harm appears to not only affect response bias, but also sensitivity, albeit to a smaller degree. The observer gender sensitivity difference persists despite conditions that should have reduced threat, consistent with a selection bias account.

Figure 7

Effect of Panel Decision (Upper Panel), Its Interaction with Employee Gender (Middle Panel) in Response Bias, and Gender Difference in Sensitivity (Lower Panel) Across Studies



Note. Error bars show 95% confidence intervals. Positive values indicate more liberal response bias for harmful than beneficial outcomes (upper panel), an interaction in the stereotype-consistent direction (middle panel), and greater sensitivity of male than female participants (lower panel). The judgment bias condition is abbreviated as "JBT". When comparing effect sizes between panels, take into account that the vertical axis is scaled differently in the different panels.

Study 4

Study 4 examines a further boundary condition of effects identified in the preceding studies: the role of the discrimination context itself. In the preceding studies, the setting of potentially discriminatory corporate pay raise decisions evokes prototypical expectation about gender discrimination and on how it should manifest in these decisions. Study 4 examined the extent to which the effects on response bias and the gender difference in sensitivity depend on the discriminatory context implemented in the Bias Detection Task.

To this end, we contrasted the stereotypical bias condition from Study 2 with a new condition, the placement condition. In the placement condition, the panel assigns employees to one of two departments based on their relative aptitude, rather than making pay raise decisions. Both departments were described as equally well-respected within the company. Consequently, and in contrast to a pay raise decision, neither type of placement decision was more harmful than the other. The stimulus design otherwise followed the same structure as the stereotypical bias condition, with the same contingency between employee gender and panel decision, but with placement rather than pay raise decisions: For decisions that do not respect the standard practiced in the learning phase, male and female employees are erroneously assigned to departments that do not correspond with their aptitude.

This placement condition thus allows us to examine the contribution of the discrimination context within the previously-found pattern of effects. First, as the placement condition is intended to remove the differential harm between the two types of decisions, we expected the effect of panel decision to be reduced in the placement relative to the stereotypical bias condition. Second, with respect to the interaction between employee gender and committee decision, the placement condition retained the contingency between employee gender and panel decision, but removed the influence of prototypes on victims of discrimination. Thus, as in Studies 1 and 2, a contingency-based learning process could pick up the implemented contingencies and thereby shape response bias, but no prototype suggests the plausibility of such a contingency. Accordingly, we expected the interaction to be reduced in the placement condition. Finally, regarding the observer difference in sensitivity, the placement condition provides a more stringent test of the threat hypothesis. Although Study 2 reversed the direction of gender bias, it still used a corporate pay raise setting that may have activated knowledge of the gender pay gap and thereby sustained differential threat for female participants. In the placement condition, erroneous placements affect male and female employees with equal frequency and are described as equally harmful, offering female participants little reason to feel more threatened than male participants. According to the threat hypothesis, the sensitivity difference between male and female participants should be reduced in the placement condition.

Methods

Participants

As preregistered, the between-participants conditions of Studies 2, 4 and 5 (stereotypical bias, reversed bias, placement, judgment bias) were sampled in one data collection with random assignment of participants to the four conditions, and in Study 4, the stereotypical bias condition already considered in Study 2 is contrasted with the placement condition. 246 people participated in the placement condition. Four of those participants were excluded because they reported a different ethnicity than White, and two 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 in the placement condition was $M = 35.78$ years ($SD = 7.82$), ranging from 18 to 49 years.

Procedure

The procedure of the placement condition was designed to be as similar as possible to the stereotypical bias condition. The participants were instructed that they would have to evaluate a series of decisions regarding the *placement* of trainee employees in one of two departments. They were told that the departments had similar reputations and that employees were paid the same wage. However, employees may have different aptitudes for the two departments and placing them in the correct department would lead to better performance and higher job satisfaction. The scores on the three criteria now indicated the relative fit to the two departments labeled "D.E." and "G.R." (see Figure 8). After the learning phase, participants were informed that there had been complaints about a number of *erroneous* placement decisions and that their task was to flag decisions that appeared erroneous for further in-depth scrutiny. The comprehension checks were analogous to the Stereotypical Bias Condition, but one comprehension check additionally tested whether participants understood the instruction on the equal reputation of the two departments had been understood. After the evaluation phase, participants answered the same self-report questions as the participants in the Stereotypical Bias Condition.

Importantly, the stimulus design followed the same design as shown for the stereotypical bias condition in Table 2. The factor "panel decision" coded whether an employee was placed in the D.E. or in the G.R. department, and the factor "merit" was replaced by relative aptitude (greater for the D.E. department versus greater for the G.R. department). Consequently, the stimulus contingencies built into the design were analogous to those implemented in the stereotypical bias condition (albeit not in line with any culturally shared stereotype).

Results

Overall sensitivity in the placement condition was 2.18 ($z = 48.1$, $p < .001$) and overall response bias was 0.183 ($z = 10.9$, $p < .001$).

Response Bias

As can be seen in Figure 9, the effect of panel decision on response bias was considerably reduced in the placement condition compared to the stereotypical bias condition. This difference in the effect between the conditions was statistically significant in our preregistered analysis ($\chi^2(1) = 4.96$, $p = .026$, $\Delta = 0.239$, 95% CI $[0.030, 0.448]$) and in the placement condition, the effect of committee decision did not differ significantly from zero ($z = 0.046$, $p = .964$, $\Delta = 0.0017$, 95% CI $[-0.153, 0.146]$). This pattern of results suggests that our instructions were effective in conveying that placements in departments D.E. versus G.R. were equally desirable, such that both types of erroneous decisions were equally harmful – in contrast to the pay raises.

Figure 8

Example Trials of the Evaluation Phase in the Stereotypical Bias (Upper Panel) and Placement Condition (Lower Panel)

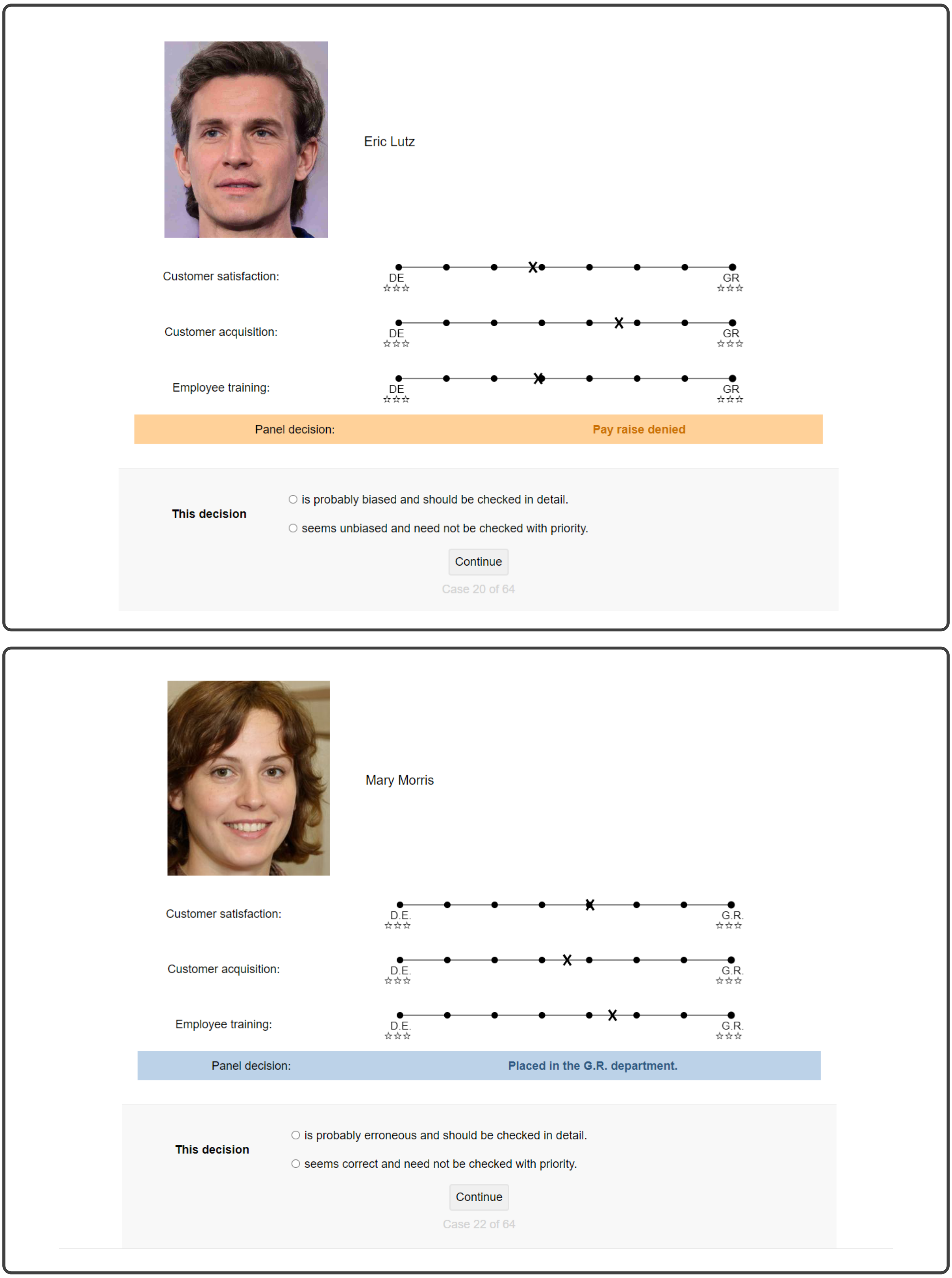


Figure 9

Panel Decision × Employee Gender Interaction in Response Bias in Study 4 (Placement Condition)

Note. Large values of response bias correspond to conservative biases, that is, less of an inclination to consider a given panel decision “biased”. Conversely, small values indicate liberal response biases. Conditions of Study 4 are abbreviated as "Stereotypical" (stereotypical bias condition) and "Placement" (placement condition). Error bars show 95% confidence intervals.

As also shown in Figure 9, the interaction between panel decision and employee gender is strongly reduced in the placement condition. This impression is supported by the preregistered analyses, which indicate that the interaction was significantly smaller in the placement condition than in the stereotypical bias condition ($\chi^2(1) = 11.01$, $p < .001$, $\Delta = 0.077$, 95% CI $[0.032, 0.123]$). In fact, the interaction between panel decision and employee gender is not statistically significantly different from zero in the placement condition ($z = 1.41$, $p = .158$, $\Delta = 0.024$, 95% CI $[-0.009, 0.057]$).

Sensitivity

Male participants were once again somewhat more sensitive than female participants in Study 4. The magnitude of this gender difference was not significantly different in the placement compared to the stereotypical bias condition ($\chi^2(1) = 0.162$, $p = .687$). As in the previous studies, it was significantly different from zero in the placement condition ($z = 2.19, p = .029$, $\Delta = 0.198$, 95% CI $[0.021, 0.376]$) and similar in size to the other conditions (see Figure 7).

Ratings

As can be seen in Tables 3 and 4, ratings of gender bias, discrimination, favoritism, and harm (Q5 – Q8) were generally much closer to the scale midpoint (2.0, indicating the absence of systematic gender bias) in the placement condition than in the stereotypical bias condition. The differences between the two conditions were significant for each of these four ratings (smallest $t = 13.12$, largest $p < .001$, smallest $d = 1.19$, 95% CI $[0.995, 1.38]$).

Furthermore, evidence emerged that participants picked up on the implemented stimulus contingencies: they reported lower frequency ratings for trial types that did not occur (Q1 and Q4) compared to trials that did occur (Q2 and Q3; $t(244) = 19.00$, $p < .001$, $d = 1.21$, 95% CI $[0.940, 1.49]$, and $t(239) = 2.43$, $p = .016$, $d = 0.157$, 95% CI $[-0.098, 0.411]$) in the stereotypical bias condition and the placement condition, respectively. However, this learning of stimulus contingencies was much stronger in the stereotypical bias condition than in the placement condition ($t(393.09) = 15.16$, $p < .001$, $d = 1.37$, 95% CI $[1.17, 1.57]$).

Discussion

The response-bias effect of panel decision seen in the stereotypical bias condition was absent in the placement condition, which indicates that participants considered the two decision outcomes equivalent. This conclusion dovetails with the rating results, such that participants perceived much less gender bias in the placement condition than in the stereotypical bias condition.

Regarding the interaction between employee gender and panel decision, the results of Study 4 indicate that although the stimulus contingencies were retained in the placement condition, they did not significantly affect participants' response bias. This outcome suggests that the contingency-based component needs to be guided by culturally shared prototypes to become effective. Prototypes may guide attention to critical patterns of panel decisions and thus support learning of actual contingencies – which in turn enables contingency-driven effects on response bias. Again, this analysis aligns with the rating data, such that explicit stimulus contingency learning was much more pronounced in the stereotypical bias condition than in the placement condition.

Finally, regarding the observer gender difference in sensitivity, the placement condition removed both the difference in harm and the discrimination context, which should reduce threat for female participants (compared to the Bias Detection Task). However, male participants again exhibited somewhat higher sensitivity than female participants in the placement condition, and the effect did not differ significantly from the stereotypical bias condition. The threat hypothesis thus again finds no support, and the selection-bias hypothesis remains the most consistent explanation of this persistent effect.

Study 5

Finally, in Study 5, we focused on disentangling the causal origin of effects in the Bias Detection Task. The pattern of results that has thus far emerged across Studies 1 to 3 may reflect differences in standard application or in attributional processing – the two components of attributions to discrimination according to the definition by Major et al. (2002, see Figure 2). (1) In standard application, observers determine whether the target individual's treatment was just or not. (2) In attributional processing, if they judged the treatment to be unjust they decide whether it was based on the target individual's social identity.

Study 5 contrasted the stereotypical bias condition from Studies 2 and 4 with a condition motivated by Axt and colleagues' (2018) Judgment Bias Task. This judgment bias condition removes the panel decisions, and the participants' task is to make pay raise decisions themselves, as in the learning phase of the Bias Detection Task. As explained in the Introduction, the judgment bias condition allows us to study the standard application stage in Major et al.'s (2002) conceptualization of attributions to discrimination in isolation. This enables us to assess which of our effects originate during standard application and which reflect attributional processing.

To illustrate, consider the consistently-observed difference in response bias between "denied" and "granted" decisions. A natural interpretation of this effect is that the harm caused by the observed panel decision affects observers' inclination to attribute the given panel decision to discrimination – reflecting attributional processing, and as already reviewed, this is consistent with the interpretations found in the literature. Alternatively, however, this effect may reflect a general tendency in standard application to consider employees as deserving pay raises (i.e., a beneficial treatment) rather than not deserving them (i.e., deserving a more harmful treatment) – independent of the presence of an actual panel decision. If such a tendency is already visible in the judgment bias condition in the form of a bias toward more favorable (i.e., grant) decisions, this would indicate that the effect observed in the Bias Detection Task already arises in the standard application stage (as captured with the judgment bias condition), and not only in attributional processing: Participants' own favorable standard would conflict more often with a "denied" decision than a "granted" one, leading them to more frequently judge "denied" decisions as biased. This rationale suggests a more liberal response bias for denied cases, which is precisely the result we observed in the Bias Detection Task in the previous studies.

A similar reasoning applies to the interaction between panel decision and employee gender on response bias. An interpretation in terms of attributional processing, which we invoked throughout this manuscript, explains this effects through expectancy- and contingency-based shifts in response bias. According to this explanation, participants would use the actual panel decision in conjunction with the employee's gender to adjust their inclination to judge the panel 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 application that might occur independently of the actual panel decision and could account for the observed interaction. This account is somewhat surprising, as it is natural to interpret this interaction in attributional terms, as we have done thus far. Nevertheless, it is not implausible, given that Axt et al., 2019 found a pro-women bias in the Judgment Bias Task.

Regarding the sensitivity difference between male and female participants we found in all conditions with the Bias Detection Task, the judgment bias condition again allowed us to test the threat hypothesis: In this condition, there is no mention of bias in panel decisions, nor do participants see any panel decisions in the evaluation phase. The judgment bias condition might therefore again be associated with little threat, which would predict a reduced sensitivity difference.

Methods

Participants

As preregistered, the between-participants conditions of Studies 2, 4 and 5 (stereotypical bias, reversed bias, placement, judgment bias) were sampled in one data collection with random assignment of participants to the four conditions, and in Study 5, the stereotypical bias condition already considered in Studies 2 and 4 is contrasted with the judgment bias condition.

246 people participated in the judgment bias condition. Three of those participants were excluded because they reported a different ethnicity than White, resulting in a total sample size of $N = 243$ (122 women). The mean age of the sample was $M = 35.14$ years ($SD = 7.93$), ranging from 18 to 49 years.

Procedure

The judgment bias condition differed from the stereotypical bias condition in that participants received no instruction thatdecisions were possibly biased. Instead, they were told to continue with the task from the learning phase (i.e., *making* pay raise decisions) and that they would no longer see a typical decision upon responding (as in the learning phase). As in the Bias Detection Task, after the learning phase the presented cases also showed demographic information. Because there was no gender bias to observe, there were no self-report questions at the end of the session in the judgment bias condition.

Data Analysis

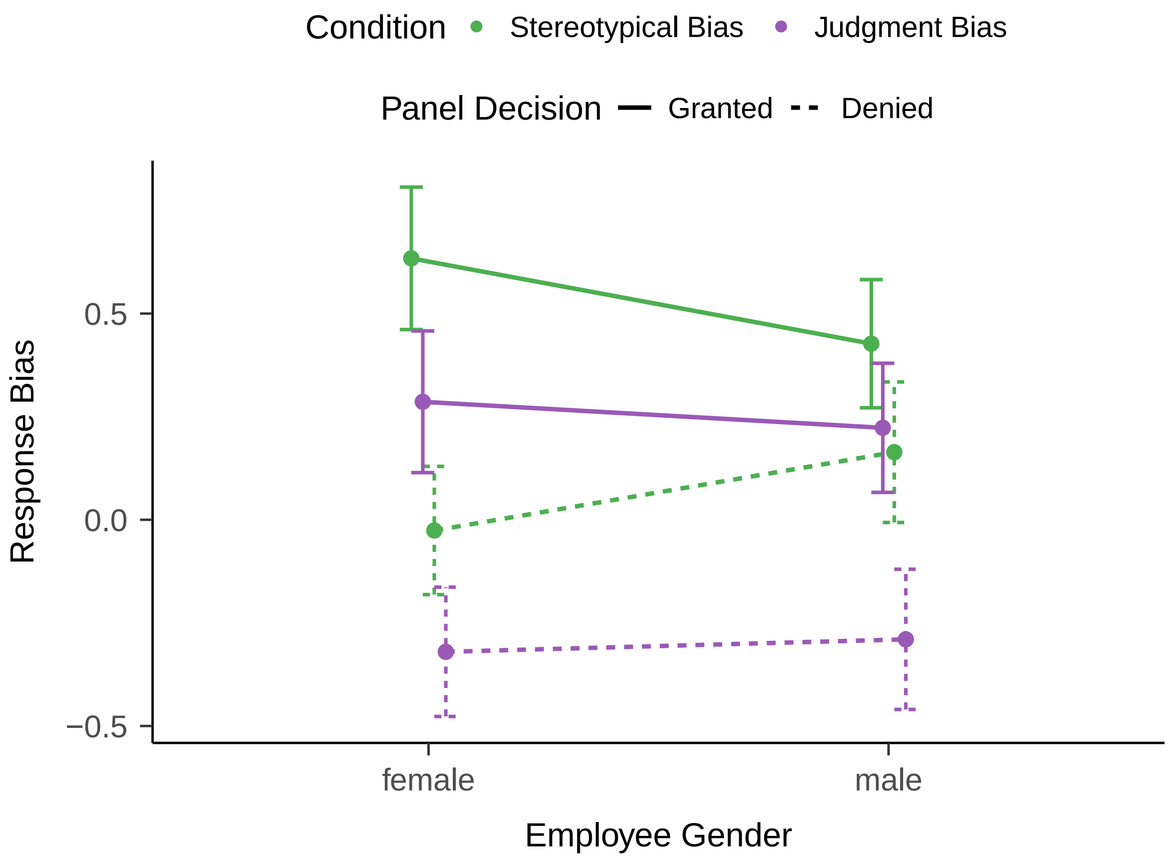
To allow us to analyze the judgment bias condition using the same metric and model as in the other conditions, a panel decision was assigned to each trial (but not shown to the participants) with the same contingencies as in the stereotypical bias condition. Participants' responses were coded as "likely biased" if they mismatched the panel decision and as "unbiased" if they matched. Consequently, the model parameters of the judgment bias condition are subject to a somewhat different interpretation than in the stereotypical bias condition with the Bias Detection Task. For instance, assume that participants in the judgment bias condition exhibit a tendency to grant rather than deny pay raises. As explained above, such a tendency would lead to more mismatches between participants' own pay raise decisions and the (internally assigned) panel decision when a pay raise was denied. This would manifest as an effect of panel decision on response bias (a bias to more frequently consider "denied" decisions as biased), similar to the corresponding effect in the Bias Detection Task.

Results

Overall sensitivity in the judgment bias condition was 1.89 ($z = 41.9$, $p < .001$) and overall response bias was -0.013 ($z = 0.786$, $p = .432$).

Figure 10

Panel Decision × Employee Gender Interaction in Response Bias in Study 5 (Judgment Bias Condition)



Note. Large values of response bias correspond to conservative biases, that is, less of an inclination to consider a given panel decision “biased”. Conversely, small values indicate liberal response biases. Conditions of Study 5 are abbreviated as "Stereotypical" (stereotypical bias condition) and "Judgment Bias" (judgment bias condition). Error bars show 95% confidence intervals.

Response Bias

As Figure 10 shows, the main effect of panel decision on response bias was similar in the judgment bias condition as in the stereotypical bias condition. The preregistered analysis confirms this pattern, indicating that the effect of panel decision did not differ significantly between the two conditions ($\chi^2(1) = 0.223$, $p = .637$, $\Delta = -0.05$, 95% CI $[-0.258, 0.159]$):

As in the stereotypical bias condition, the effect was also significantly different from zero in the judgment bias condition ($z = 3.80$, $p < .001$, $\Delta = 0.286$, 95%CI $[0.138, 0.433]$).

Thus, participants in the judgment bias condition exhibited a tendency to grant rather than deny pay raises to employees.

In contrast, as also shown in Figure 7, the size of the interaction of panel decision and employee gender in response bias is much reduced in the judgment bias condition ($\chi^2(1) = 12.58$, $p < .001$, $\Delta = 0.077$, 95% CI $[0.035, 0.120]$). In fact, the interaction was not significantly different from zero in the judgment bias condition, $z = 1.58$, $p = .114$, $\Delta = 0.0238$, 95%CI $[-0.0058, 0.0534]$. This pattern of results indicates that participants in the judgment bias condition did not consider employees of either gender to be more deserving of a pay raise.

Sensitivity

Male participants were once again somewhat more sensitive than female participants in Study 5. The magnitude of this gender difference did not differ significantly between the judgment bias condition and the stereotypical bias condition, $\chi^2(1) = 2.20$, $p = .138$, $\Delta = 0.189$, 95% CI $[-0.060, 0.438]$. This result once again contradicts the threat hypothesis as an explanation of this consistent effect.

Discussion

Our results revealed a differentiated pattern: The effect of panel decision on response bias was not significantly reduced in the judgment bias compared to the stereotypical bias condition. This finding indicates that the effect originates already in the standard application stage, meaning that participants themselves have an overall response bias to grant pay raises rather than deny them when in doubt. This bias in the application of a standard (as assessed in the judgment bias condition) implies that “pay raise denied” decisions are more likely to be seen as unjust in standard application and, all other things being equal, to be attributed to discrimination than “pay raise granted” decisions (see Figure 2). In other words, the response-bias effects of harm found in attributions to discrimination across the present studies may not purely reflect effects of harm on attributional processes. Instead, these effects may reflect an overall tendency towards magnanimity in determining the appropriate action.

This pattern is qualitatively different for the interaction of panel decision and employee gender: It occurred reliably in the present studies using the Bias Detection Task, but was absent from the judgment bias condition. This pattern of results therefore underscores our argument that its origin lies in the stage of attributional processing, which it shapes through culturally shared prototypes of discrimination (via prototype-based response bias) and learned contingencies (via contingency-based response bias).

Study 5 thus yields a clear and differentiated answer to the question of causal origin: the effect of harm on response bias appears to originate in standard application, reflecting a general tendency to grant pay raises when the evidence is ambiguous. In contrast, the interaction between harm and employee gender originates in attributional processing, shaped by culturally shared prototypes and learned contingencies. Notably, although these two effects share a common mediating mechanism — response bias — they differ fundamentally in their causal origin.

General Discussion

SDT has been a popular theoretical lens for conceptualizing attributions to discrimination. However, to date, no research has explored the potential of SDT to empirically clarify the mediation of observed effects through sensitivity and response bias. The present research takes a first step towards filling this empirical and theoretical gap. As an initial step, we examined gender bias in pay raise decisions and developed an experimental paradigm, the Bias Detection Task, that enables estimation of the SDT parameters response bias and sensitivity in attributions to discrimination. The present research thereby contributes to our theoretical understanding of attributions to discriminations in two ways: First, we used the Bias Detection Task to decompose central effects studied in this literature – notably, of harm, victim prototypicality, contingencies and observer gender differences – according to their causal pathway, distinguishing effects on response bias from effects on sensitivity. Second, we decomposed effects according to their causal origin, building on Major et al.'s (2002) two-components conceptualization of attributions to discrimination according to which effects can originate in standard application or in subsequent attributional processing. In the following, we first summarize our main findings before turning to limitations, future directions and implications.

Harm Affects Response Bias More Than Sensitivity

Harm, as manipulated by panel decision, exerted a considerable effect on observers' response bias: Observers are more inclined to attribute harmful decisions to discrimination than beneficial ones, regardless of evidence of discrimination. This finding is in line with arguments by Swim et al. (2003) who found that harmful consequences increase attributions to discrimination even in the absence of clear cues to discrimination. However, Barrett and Swim (1998) instead related harm to signal intensity, speculating that it might increase sensitivity. We did find an effect of harm on sensitivity in Study 3, which realized the largest statistical power, and also in a joint analysis of all conditions of the Bias Detection Task. That said, effects on sensitivity were much smaller (0.07 units on the latent strength-of-evidence scale in the joint analysis) than the effects on response bias (between 0.24 and 0.47 on that scale).

With respect to the distinction between standard application and attributional processing, the effect of harm on response bias could originate in either or both decision stages: It may reflect a bias in standard application, that is, in determining whether the employee in question deserves a pay raise or not. Alternatively, as the conventional interpretation would suggest, it may also be caused by a bias in attributional processing, whereby disfavoring decisions are attributed to discrimination with greater likelihood than favoring decisions. The results of Study 5 clearly place the effect in the standard application stage rather than in the attribution processing stage; it reflects observers' bias to grant rather than deny a pay raise when in doubt. More generally, the present results are consistent with the novel interpretation that similar effects of harm reported in the wider literature may not purely reflect effects on attributional processes per se, but may, at least partly, reflect an overall tendency towards magnanimity in determining a just decision.

Harm and Victim Gender Interact in Response Bias, but Not in Sensitivity, Through Contingency-Based and Prototype-Based Components

In addition, a robust two-way interaction consistently emerged between panel decision and employee gender in response bias. Across our experiments, we produced strong evidence for a contingency-based component and a prototype-based component jointly shaping this interaction. The contingency-based component reflects learning of contingencies through exposure to the actual decisions, whereas the prototype-based component reflects responses under uncertainty guided by culturally shared prototypes on how gender discrimination manifests in corporate pay raise decisions. The mediation of victim prototypicality effects through response bias is in line with theoretical arguments and speculation by Heineck and Deutsch (2025) and Inman and Baron (1996).

Moreover, the contingency-based component depends on the presence of a prototype-based component: Contingencies only affected observers' judgments when existing prototypes suggested that certain contingencies might be present, as found in Study 4. This pattern suggests that prototypical expectations guided attention to critical patterns in observed decisions, enabling the learning process underlying the contingency-based contribution.

This attentional guidance by prototypes also adds plausibility to the idea that prototype-consistent decisions (women being denied a pay raise and men being granted one) might receive especially attentive and careful scrutiny, leading to an interaction between panel decision and employee gender in sensitivity. Study 3 was designed to permit a test of this interaction, but found no evidence for it, contrary to Inman's (2001) speculation about prototypical expectations shaping sensitivity.

Turning to the decomposition in terms of standard application and attributional processing, the interaction could be caused by participants holding male versus female employees to different standards in standard application. Alternatively, the interaction could be caused by participants' contingency-based or prototype-based beliefs about which cases are likely biased – beliefs that guide and shape attributional processing. Contrary to the main effect of harm, this interaction clearly has its origin in the stage of attributional processing rather than in standard application and thus, attributional processing reflects culturally-shared prototypes of discrimination (via prototype-based response bias) and learned contingencies (via contingency-based response bias).

Observer Gender Differences in Response Bias and Sensitivity?

In the present studies, only mixed evidence emerged for the ideas reviewed in the Introduction that women, as members of the lower-status group, might have a more liberal response bias and/or stronger sensitivity than men. The expected gender difference in response bias appeared in Study 3, but not in Study 1 and 2. In a joint analysis of Studies 1 to 3, the effect was also significant and not moderated by condition.

Regarding sensitivity, men were consistently found to be somewhat more sensitive than women. We ruled out explanations for this unexpected finding in terms of speed-accuracy trade-offs and differential threat experienced by men and women. Furthermore, demographic differences between the male and female samples did not contribute to this effect (see the Online Appendix).

An additional plausible explanation raised by reviewers was that (White) men, as prototypical perpetrators in the context of discrimination, may be particularly vigilant for cues for discrimination out of concern about appearing biased themselves, which could increase sensitivity. Although this is a plausible possibility, it encounters similar difficulties as the speed-accuracy hypothesis and the explanation in terms of differential threat: Increased vigilance and caution in male participants would suggest longer response times. However, as exemplified in Study 1, male participants' response times were descriptively *faster* than those of female participants (see Study 1). This difference was significant in a joint response time analysis of all studies ($F(1, 1694) = 13.61$, $p < .001$, $\eta_p^2 = .008$; see the Online Appendix for details), strongly contradicting the speed-accuracy hypothesis. In addition, this sensitivity difference is not significantly reduced in the placement condition, which implemented a context where male participants are no more prototypical perpetrators than female participants. Thus, although the present pattern of results does not rule out this account, it does reduce its plausibility.

Another possibility is that the participant gender difference in sensitivity indicates a genuine population difference in women's and men's ability to accurately judge cases in the Bias Detection Task (regardless of context and threat), reflecting gender differences in cognitive processes required for this task. Although such gender differences in higher cognition tend to be small and relatively rare (Hyde, 2014), research suggests a robust advantage for men in visuospatial skills (Hyde, 2014; Kheloui et al., 2023; Linn & Petersen, 1985). Because the employees' scores in the Bias Detection Task were visually indicated by spatial position (see, for example, Figure 8), visuospatial skills may be required to integrate employees' scores in the Bias Detection Task. Our results neither contradict nor confirm such an account. However, as elaborated throughout this manuscript, the sensitivity difference may instead plausibly reflect selection bias, which is why we are hesitant to interpret this finding in terms of a true difference between the populations of men and women. The selection-bias account also suggests caution in interpreting the response-bias difference between men and women.

More broadly, our exploration of explanations for the sensitivity difference underlines cautionary remarks from the methodological literature (e.g., Shadish et al., 2002) regarding the interpretation of main effects of participants' social group membership: Common recruitment procedures, especially self-selection, can introduce selection biases that are generally difficult to rule out and may provide plausible alternative explanations for such effects.

Limitations and Constraints on Generality

The present research contributes to theoretical analyses of attributions to discrimination by introducing two decompositions: One based on SDT, distinguishing between sensitivity and response bias (based on Barrett & Swims, 1998, influential conceptualization), and the other one in terms of processing stages, standard application and attributional processing (based on Major et al., 2002). Like any causal analysis, both decompositions rest on assumptions that can be questioned and that future work might examine critically. As a mathematical model, SDT relies on auxiliary distributional assumptions that may not be met – in this case, the prevalent assumption of Gaussian signal and noise distributions with equal variance. To ensure that our pattern of results is not contingent on this common, yet arbitrary assumption, we conducted robustness analyses with alternative distributional assumptions (i.e., logistic, Gumbel-Min or Gumbel-Max evidence distributions). Our pattern of results – that is, the effects of harm, victim prototypicality and the reported moderation through the experimental conditions on response bias and the consistent participant gender difference in sensitivity – remained consistent across all implemented distributions (see the Online Appendix for a more detailed description of these analyses). These convergent findings substantially increase our confidence that our conclusions reflect genuine psychological phenomena rather than artifacts of modeling choices. Nevertheless, further validation of the model as a measurement tool would strengthen the validity of specific (distributional) assumptions and obviate the requirement of robustness across different assumptions (Klauer, 2024).

The second decomposition follows a time-honored subtraction logic, pioneered by Donders (1868). According to this logic, we can "subtract" the outcome of the judgment bias task - which assesses standard application – from the outcome of the Bias Detection Task – which assesses the combination of standard application and attributional processing – to isolate the contribution by attributional processing. This rests on the assumption of "pure insertion" (e.g., Ulrich et al., 1999), namely that an additional process (i.e., attributional processing) can be combined with another process (i.e., standard application) in a new task without altering the first process. Again, this assumption can be questioned. Therefore, we interpreted the results cautiously, as suggestive of novel alternative interpretations, rather than conclusive evidence, of effects that are conventionally construed as effects on attributional processing. More generally, this analysis highlights the role of differential standards which, in addition to attributional processes, may contribute to observed effects in attributions to discrimination.

Regarding the generality of our findings, it is important to note that the present work was only a first step towards unlocking the full potential of SDT for studying attributions to discrimination. Consequently, the generality of the present research is also limited in many ways. First, we examined only one kind of bias within a specific context: gender bias in corporate pay raise (and placement) decisions. We also held ethnicity constant by presenting only portraits of White people to a sample of only White participants. These decisions partly reflect our desire to enhance statistical and internal validity (Shadish et al., 2002) by limiting the heterogeneity of the materials and participant sample. These methodological decisions necessarily restrict the generalizability of our findings. At the same time, they also reflect our belief that generalizability may be limited for factual rather than only methodological reasons: The pattern of effects might be different in different arenas. For example, differences in perceptions of sexism between men and women are often small (Barreto & Ellemers, 2015), whereas White and Black people perceive racism differently (for a review, see Carter and Murphy, 2015).

Furthermore, evaluative contexts can differ substantially in the kind of available information, the ambiguity of decision criteria, and the perceived stakes of outcomes. For example, pay raise decisions are often based primarily on evaluations of past performance, whereas hiring decisions or university admissions are more strongly oriented toward predicting future success and potential. These differences may involve partly distinct psychological processes and may therefore map differently onto sensitivity and response bias. As a result, attributions to bias in pay raise decisions may not necessarily generalize straightforwardly to other contexts. For such reasons, we believe that attributions to gender bias in ethnically heterogeneous samples involving employees of different ethnicities, attributions to racial bias, and intersectional effects, as well as applications to other evaluative and interpersonal contexts deserve additional SDT studies in their own right. These extensions, along with several others that we highlight below, open a broad agenda for future research with the Bias Detection Task.

Future Research Directions

Elaborating on this line of thought, future research might examine how the structure of the decision context itself shapes sensitivity and response bias. In the present studies, participants evaluated corporate pay raise and placement decisions in contexts in which positive outcomes were common. However, many real-world evaluative settings —- such as highly selective university admissions, hiring decisions, or competitive scholarship allocations – involve much lower base rates of positive outcomes, which may alter assumptions about what constitutes a fair decision and produce different response tendencies.

Similarly, future studies could investigate the role of prior beliefs about discrimination within a given institution or context. Participants may not only rely on the information presented in the task itself, but also on assumptions about how likely a particular company, organization, or social environment is to discriminate (e.g., Heineck & Deutsch, 2024; O'Brien et al., 2008; Ruggiero & Taylor, 1997). Such beliefs could influence response bias independently of participants' actual ability to distinguish discriminatory from non-discriminatory decisions. Manipulating organizational reputation or providing information about the prevalence of past discrimination may therefore help clarify how contextual expectations shape attributions to discrimination.

It is also possible to manipulate the history of past discrimination at the trial level, as an alternative or additional way to provide cues for possible discrimination: In everyday life and in studies of attributions to discrimination, cues to possible discrimination often involve decisions or treatments that are unfair. Alternatively (or additionally), cues to possible discrimination could involve information on a history of biased decisions. For instance, the Bias Detection Task could be modified such that the presented cases show decisions by different decision makers, whose history of prior decisions is presented to participants (i.e., information equivalent to that tabulated in Table 2). Noise trials would present a decision history showing no contingency on gender (in studies of gender discrimination), whereas signal trials would present a history showing such a contingency. The strength of this evidence could be manipulated via the number of prior decisions available for a given decision maker, the magnitude of that decision maker's contingency, or both.

A different question concerns the malleability of observers' sensitivity. In the present studies, our manipulations more readily shifted participants' response bias than their ability to accurately discriminate biased from unbiased decisions. This pattern may partly reflect characteristics of the task and the incentives under which participants operated. Future research could therefore examine the boundary conditions of this finding – for example, whether it remains robust when accurate judgments are rewarded, accountability pressures are introduced, or lay participants are compared with trained professionals whose real-world roles depend on accurate evaluations.

Finally, the Bias Detection Task also offers a fitting framework for studying individual differences in attributions to discrimination. Variables such as self-esteem (Major et al., 2003) and ideological beliefs about the prevalence and severity of discrimination (Major et al., 2007; see also Miller and Saucier, 2018) have been shown to be associated with the frequency with which individuals attribute to discrimination. The Bias Detection Task can help determine whether such associations reflect individual differences in response bias, sensitivity, or both.

Theoretical and Practical Implications

Based on the discussion of the present findings, we can return to the potential theoretical and practical implications proposed in the introduction and address them in light of the evidence. Regarding the differentiation between response bias and sensitivity, across all studies we found response bias to be more malleable than sensitivity, with stronger and more diverse effects for the former compared to the latter.

This finding is perhaps unsurprising, given that response biases are partly shaped by observers' response strategies, expectations and goals which are, to some extent, under the observers' control. In contrast, sensitivity turned out to be less strongly affected and, as a measure of ability, may generally be less malleable.

For applied efforts to improve the detection of discrimination, this pattern of results is both good and bad news. On the one hand, response bias is responsible for systematic biases in attributions to discrimination and is also likely to be malleable under remedial training and awareness-raising initiatives. However, on the other hand, the silver bullet in improving the detection of discrimination is not adapting response bias, but rather increasing sensitivity. Increasing sensitivity simultaneously reduces both types of possible errors, misses, and false alarms. In contrast, changes in response bias always decrease one type of error at the expense of increasing the other (see Figure 1).

Returning to theoretical implications, our findings regarding the decision stages in attributions to discrimination revealed a more differentiated pattern: the effect of harm originates in standard application, whereas the interaction between harm and employee gender emerges in attributional processing. These results underline the sometimes overlooked role of the first component of Major et al.'s (2002) two-components view of attribution to discrimination: Standard application — the determination of what a just or fair decision would be in the observed context — can be the source of downstream effects on attributions to discrimination, rather than attributional processing proper. Taken together, we hope that the Bias Detection Task and the analytical framework demonstrated here will prove useful for future research aimed at understanding, and ultimately improving the detection of discrimination.

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.

Axt, J. R., Casola, G., & Nosek, B. A. (2019). Reducing social judgment biases may require identifying the potential source of bias. *Personality and Social Psychology Bulletin*, *45*(8), 1232–1251.

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. (2015). Detecting and experiencing prejudice. In *Advances in Experimental Social Psychology* (pp. 139–219, Vol. 52). Elsevier. https://doi.org/10.1016/bs.aesp.2015.02.001

Barrett, L. F., & Swim, J. K. (1998). Appraisals of prejudice and discrimination. In J. K. Swim & C. Stangor (Eds.), *Prejudice: The target's perspective* (pp. 11–36). Academic Press. https://linkinghub.elsevier.com/retrieve/pii/B9780126791303500363

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

Branscombe, N. R., Schmitt, M. T., & Harvey, R. D. (1999). Perceiving pervasive discrimination among African Americans: Implications for group identification and well-being. *Journal of Personality and Social Psychology*, *77*(1), 135–149. https://doi.org/10.1037/0022-3514.77.1.135

Carlsson, R., & Sinclair, S. (2018). Prototypes and same-gender bias in perceptions of hiring discrimination. *The Journal of Social Psychology*, *158*(3), 285–297. https://doi.org/10.1080/00224545.2017.1341374

Carter, E. R., & Murphy, M. C. (2015). Group-based differences in perceptions of racism: What counts, to whom, and why? *Social and Personality Psychology Compass*, *9*(6), 269–280. https://doi.org/10.1111/spc3.12181

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

Doering, L., Doering, J., & Tilcsik, A. (2023). "Was It Me or Was It Gender Discrimination?" How Women Respond to Ambiguous Incidents at Work. *Sociological Science*, *10*, 501–533. https://doi.org/10.15195/v10.a18

Donders, F. C. (1868). Die schnelligkeit psychischer processe: Erster artikel. *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*, 657–681.

Harris, M. M., Lievens, F., & Van Hoye, G. (2004). "I Think They Discriminated Against Me": Using Prototype Theory and Organizational Justice Theory for Understanding Perceived Discrimination in Selection and Promotion Situations. *International Journal of Selection and Assessment*, *12*(1-2), 54–65. https://doi.org/10.1111/j.0965-075X.2004.00263.x

Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology*, *37*(1), 25–38. https://doi.org/10.1037/0022-3514.37.1.25

Heineck, P.-M., & Deutsch, R. (2024). Summarized and sequential discrimination - A paradigm for research on the perception of multiple instances of discrimination. *Journal of Experimental Social Psychology*, *110*, 104548. https://doi.org/10.1016/j.jesp.2023.104548

Heineck, P.-M., & Deutsch, R. (2025). More likely or more wrong? - Disentangling the prototype effect of discrimination perception. *Journal of Experimental Social Psychology*, *118*, 104731. https://doi.org/10.1016/j.jesp.2025.104731

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. E. Higgins & A. Kruglanski (Eds.), *Social Psychology: Handbook of Basic Principles*. Guilford.

Hyde, J. S. (2014). Gender similarities and differences. *Annual review of psychology*, *65*, 373–398.

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

Inman, M. L., Huerta, J., & Oh, S. (1998). Perceiving discrimination: The role of prototypes and norm violation. *Social Cognition*, *16*(4), 418–450. https://doi.org/10.1521/soco.1998.16.4.418

Jakob, M., Hartmann, R., Klauer, K. C., & Meyer-Grant, C. G. (2025). The statistical costs of two-step signal detection analyses: A case for a maximum likelihood mixed-effects approach [Preprint, posted December 12, 2025]. https://doi.org/10.31234/osf.io/w59ma_v1

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

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

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

Kheloui, S., Jacmin-Park, S., Larocque, O., Kerr, P., Rossi, M., Cartier, L., & Juster, R.-P. (2023). Sex/gender differences in cognitive abilities. *Neuroscience & Biobehavioral Reviews*, *152*, 105333.

Klauer, K. C. (2024). Mathematical and computational models. In H. T. Reis, T. West, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 653–677). Cambridge University Press.

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

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child development*, 1479–1498.

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., Kaiser, C. R., O'Brien, L. T., & McCoy, S. K. (2007). Perceived discrimination as worldview threat or worldview confirmation: Implications for self-esteem. *Journal of Personality and Social Psychology*, *92*(6), 1068–1086. https://doi.org/10.1037/0022-3514.92.6.1068

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* (pp. 251–330, Vol. 34). 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

Merritt, S. M. (2026). A missed opportunity: How signal detection theory can advance research on prejudice detection. *Frontiers in Organizational Psychology*, *4*, 1629459.

Miller, S. S., & Saucier, D. A. (2018). Individual differences in the propensity to make attributions to prejudice. *Group Processes & Intergroup Relations*, *21*(2), 280–301. https://doi.org/10.1177/1368430216674342

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

Pinel, E. C. (2004). You're just saying that because I'm a woman: Stigma consciousness and attributions to discrimination. *Self and Identity*, *3*(1), 39–51. https://doi.org/10.1080/13576500342000031

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

Ruggiero, K. M., & Taylor, D. M. (1997). Why minority group members perceive or do not perceive the discrimination that confronts them: The role of self-esteem and perceived control. *Journal of Personality and Social Psychology*, *72*, 373–389. https://doi.org/10.1037/0022-3514.72.2.373

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

Schmader, T., Dennehy, T. C., & Baron, A. S. (2022). Why antibias interventions (need not) fail. *Perspectives on Psychological Science*, *17*(5), 1381–1403. https://doi.org/10.1177/17456916211057565

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, *115*(2), 336–356. https://doi.org/10.1037/0033-295X.115.2.336

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.

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

Skarlicki, D. P., & Kulik, C. T. (2004). Third-Party Reactions to Employee (Mis)Treatment: A Justice Perspective. *Research in Organizational Behavior*, *26*, 183–229. https://doi.org/10.1016/S0191-3085(04)26005-1

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

Ulrich, R., Mattes, S., & Miller, J. (1999). Donders's assumption of pure insertion: An evaluation on the basis of response dynamics. *Acta Psychologica*, *102*, 43–76. https://doi.org/10.1016/S0001-6918(99)00019-0

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