A contest study to reduce attractiveness-based discrimination in social judgment

Eliane Roy, Bastian Jaeger, Anthony M. Evans, Kate M. Turetsky, Brian O'Shea, Michael B. Peterson, Balbir Singh, Joshua Correll, Denise Zheng, Kirk Warren Brown, Erika L. Kirgios, Linda W. Chang, Edward H. Chang, Jennifer Steele, Julia Sebastien, Jennifer Sedgwick, Amy Hackney, Rachel Cook, Xin Yang, Arin Korkmaz, Jessica J. Sim, Nazia Khan, Maximilian Primbs, Gijsbert Bijlstra, Ruddy Faure, Johan C. Karremans, Luiza A. Santos, Jan G. Voelkel, Maddalena Marini, Jacqueline M. Chen, Teneille Brown, Haewon Yoon, Carey Morewedge, Irene Scopelliti, Neil Hester, Xi Shen, Ming Ma, Danila Medvedev, Emily G. Ritchie, Chieh Lu, Yen-Ping Chang, Aishwarya Kumar, Ranjavati Banerji, Jeremy D. Gretton, Landon Schnabel, Bethany Teachman, Ariella Kristal, Kao-Wei Chua, Jonathan B. Freeman, Sean Fath, Lusine Grigoryan, Isabelle M. Weißflog, Yalda Daryani, Reza Pourhosein, Stephanie Johnson, Elsa Chan, Samantha M. Stevens, Stephen Anderson, Roger Beaty, Sandro Rubichi, Veronica Margherita Cocco, Loris Vezzali, Calvin K. Lai, Jordan R. Axt

Author Contributions:
E. Roy, C.K. Lai and J.R. Axt developed the study concept. All authors contributed to the study design. Data collection and data analysis was led by E. Roy under the supervision of J.R. Axt. E. Roy and J.R. Axt drafted the manuscript, and all authors had the opportunity to provide revisions. All authors approved the final manuscript for submission.

Abstract

Discrimination in the evaluation of others is a key cause of social inequality around the world. However, relatively little is known about psychological interventions that can be used to prevent biased evaluations. The limited evidence that exists on these strategies is spread across many methods and populations, making it difficult to generate reliable best practices that can be effective across contexts. In the present work, we held a research contest to solicit interventions with the goal of reducing discrimination based on physical attractiveness using a hypothetical admissions task. Thirty interventions were tested across four rounds of data collection (total $N >$ 20,000). Using a Signal Detection Theory approach to evaluate interventions, we identified two interventions that reduced discrimination by lessening both decision noise and decision bias, while two other interventions reduced overall discrimination by only lessening noise or bias. The most effective interventions largely provided concrete strategies that directed participants' attention towards decision-relevant criteria and away from socially biasing information, though the fact that very similar interventions produced differing effects on discrimination suggests certain key characteristics that are needed for manipulations to reliably impact judgment. The effects of these four interventions on decision bias, noise, or both also replicated in a different discrimination domain, political affiliation, and generalized to populations with self-reported hiring experience. Results of the contest for decreasing attractiveness-based favoritism suggest that identifying effective routes for changing discriminatory behavior is a challenge, and that greater investment is needed to develop impactful, flexible, and scalable strategies for reducing discrimination.

Word count: 250

Statement of limitations

The studies presented in this article should be considered in light of some limitations. First, the conclusions from this work are limited by the narrow scope in which discrimination was assessed. In all studies, we used the same paradigm (the Judgment Bias Task) to measure discrimination, and thus we cannot claim to understand the effectiveness of interventions in different discrimination contexts, such as various real-world settings. In addition, we investigated discrimination in only one judgment context (admission decisions) and two domains (physical attractiveness and political affiliation), meaning results could differ if extended to other contexts or domains in which discrimination occurs. In addition, interventions tested in the studies were not subject to manipulation checks, meaning we cannot rule out the possibility that findings would change if we could ensure all participants complied with their intervention's instructions. Our choice of domains (attractiveness and political ideology) also meant that most of the interventions were adapted from research on other forms of discrimination, and we may have seen more success if using a form of discrimination (e.g., based on race) that had a richer literature of discrimination-reducing interventions. Lastly, although samples included both novice participants and those with self-reported hiring experience, these participants came primarily from the US or Canada. As such, results may differ in other populations.

Word count: 218

A contest study to reduce attractiveness-based discrimination in social judgment

People rely on social categories to navigate their environment. One problem with this tendency is that our perceptions, beliefs, and behaviors can be biased by social information, even when such information is non-diagnostic or irrelevant (Wilson & Brekke, 1994). As a result, discrimination – the prejudicial treatment of one social group over another – can occur, contributing to disparities in many spheres of life, from academia (Milkman et al., 2012; Moss-Racusin et al., 2012), to employment (Ameri et al., 2015), to policing (Hester & Gray, 2018) and other economic outcomes (Doleac & Stein, 2013; Edelman et al., 2017).

The far-reaching consequences of social category bias have inspired a wealth of research in social psychology on prejudice, stereotyping, and discrimination, with one analysis finding that around one in every eight articles from leading psychology journals focused on these topics (Dovidio & Gaertner, 2010). Yet, while the presence and impact of discrimination has been well-documented, comparatively less progress has been made in the development and implementation of scalable and generalizable interventions for reducing discrimination, especially in the context of interpersonal evaluation. In fact, many popular methods used by organizations to curb discrimination (e.g., diversity training, implicit bias workshops, multicultural education) often lack support from both theory and research (Paluck & Green, 2009; Dobbin & Kalev, 2016; Chang et al., 2019; Lai & Lisnek, 2023). In this work, we explore how a variety of potential interventions submitted by social scientists fared in lessening discrimination based on physical attractiveness, a prominent and pervasive form of favoritism in social judgment (e.g., Feingold, 1992; Lippens et al., 2023; Mobius & Rosenblat, 2006).

Prior research has identified several interventions that show promise for effectively reducing socially biased judgment across a broad array of contexts and outcomes. Some of these

approaches target decision-makers' motivation or provide evaluators with helpful strategies for navigating judgment. For instance, one study found that increasing feelings of accountability among undergraduates reduced biases in reliance on first impressions during a mock job interview (Webster et al., 1996), while other work (Mendoza et al., 2010) has shown that giving undergraduates an "if-then" judgment strategy for making decisions reduced racial bias in a task that required quickly identifying guns or harmless objects in the hands of Black or White men (e.g., "If I see a person, then I will ignore his race!").

Other interventions have looked to shift aspects of the decision-making process. For instance, one series of studies (Uhlmann & Cohen, 2005) found that asking adult participants to commit to prioritizing specific criteria before a mock hiring task eliminated gender-based discrimination. In another example, providing undergraduates with a more subjective scale for rating the value of a piece of written work (e.g., a scale from "worth very little money" to "worth lots of money") lessened reliance on the journalist's gender relative to using a more objective scale (e.g., inputting a value between 50$ and 1000$; Biernat & Manis, 1994). Alternatively, manipulating choice architecture by having decision-makers evaluate candidates jointly rather than separately eliminated gender biases in hypothetical selection decisions (Bohnet et al., 2016). Finally, initially partitioning candidates into groups (e.g., by gender, nationality or university) has led to more diverse selections without any changes to average applicant competence (Feng et al., 2020). More broadly, debiasing approaches in decision-making have shown several successful applications to the context of discriminatory judgment (see Soll et al., 2015 for review).

While the findings from these studies are useful for understanding what could reduce discrimination, they are difficult to directly compare with each other. This is because the studies

were conducted across a wide range of populations, a diverse set of procedures, and many different discrimination-related outcomes. As a result, the current literature lacks comparative evidence across different intervention strategies that use the same sample source and outcome. When studies do not share such properties, determining the relative efficacy of each intervention is challenging. In other words, when two interventions are tested under different conditions (e.g., using samples from different populations or different measures of discrimination), it is unclear the degree to which differences in the interventions' effectiveness are due to aspects of the sample or the operationalization of discrimination.

One means of addressing this issue would be for researchers to develop interventions and test them sequentially over a period of time, using the same sample source and outcome measure across studies. While this approach has the benefit of being relatively simple in execution, it is time-consuming and inefficient, as each intervention needs to be tested one after the other. Moreover, temporal shocks or historical movements occurring during the testing of the many interventions can create inconsistency in responses even across the same sample source. In addition, the breadth of interventions deployed is limited to a single research team, who likely share assumptions about what interventions will be more or less effective. This narrow thinking can reduce the possible number of interventions created or selected for testing. That is, many potentially successful intervention approaches could be excluded simply due to a lack of familiarity with certain research literatures.

In this project, we adopted a different methodology for identifying effective interventions for reducing one form of socially biased judgment: discrimination based on physical attractiveness. Specifically, we held a "research contest" to determine the most effective interventions for reducing attractiveness-based discrimination, using the same sample and

outcome measure across interventions. A contest approach allows for a greater diversity of interventions to be tested and can be completed in a relatively short timeframe, and thus provides many advantages over the sequential testing approach.

The effectiveness of these interventions was tested on a judgment task known to reliably produce discrimination that favors more over less physically attractive people (Axt et al., 2018). A great deal of prior research has documented the robust and impactful consequences of favoritism based on physical attractiveness, and such discrimination has been tied to wage disparities (Hamermesh & Biddle, 1993; Mobius & Rosenblat, 2006; Monk et al., 2021), unfair promotion to leadership positions (Nault et al., 2020), unequitable voting outcomes (Berggren et al., 2010), biased teaching evaluations (Felton et al., 2008), and unjust hiring outcomes (Lippens et al., 2023).

Psychological Insights through Research Contests

Research contests have a long history in science. One of the first psychological studies that used a contest design focused on strategies to navigate the prisoner's dilemma (Axelrod, 1980). Researchers recruited experts in game theory for a computer tournament that aimed to find the most effective way to play the prisoner's dilemma game across multiple rounds. Of the fourteen entries in the tournament, the simplest strategy, named "tit-for-tat," yielded the best results, and "tit-for-tat" has since had a marked influence in the fields of economics and psychology and has helped propel research on cooperative behavior (Kopelman, 2020).

Similar crowdsourcing approaches have been used more recently in psychological research and have shown great promise in advancing our understanding of processes like prejudice, health behavior, and social cohesion (Bruneau et al., 2018; Lai et al., 2014, 2016). For example, Bruneau et al. (2018) tested interventions to reduce people's tendencies to blame all

Muslims as a collective for the harmful actions of only a few Muslims (i.e., Muslim terrorists). In all, the intervention tournament tested eight videos that challenged participants' homogenous perception of Muslims through strategies like presenting Muslims as a diverse group, engaging with counter-stereotypical Muslims, or viewing a confrontational media interview. They found that the most effective intervention was one that made participants reflect on the hypocrisy they displayed by blaming all Muslims for the violent acts of a small group of Muslims yet failing to do the same when a small group of White people committed violence. Follow-up research then found that this intervention reduced collective blame towards Muslims for up to one year after its administration (Bruneau et al., 2020). Another recent example of a contest study aimed to identify the best method to reinforce Americans' attitudes towards democracy, further illustrating how the contest approach is being applied to a wide range of issues (Voelkel et al., 2023). The findings of these studies demonstrate that intervention tournaments can be a valuable method to efficiently identify successful strategies before moving on to more resource-intensive studies that can further dissect long-term efficacy and generalizability.

Most relevant to the present work is the intervention tournament led by Lai et al. (2014), which tested the effectiveness of 17 interventions to reduce implicit racial bias, measured by performance on an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), which measured positive and negative associations towards White and Black people. The contest identified eight interventions that reduced implicit racial bias in the short term, and these interventions were also used in a follow-up study to examine their effectiveness over time (Lai et al., 2016). Here, we adopt a similar approach but use an outcome that is more connected to discriminatory behavior, though many of the intervention approaches we test share similarities with those previously used for reducing intergroup bias in implicit associations.

In all, intervention tournaments allow researchers to compare many different ideas to a single control condition, with one defined goal: finding what works (Hameiri & Moore-Berg, 2022). Adopting a contest approach can accelerate progress on practical and theoretical issues inherent in reducing discriminatory behavior. From a practical perspective, it is possible that any intervention identified as able to reduce discrimination in a research contest could be incorporated into real-world field studies of biased behavior (e.g., Chang et al., 2019). From a theoretical perspective, researchers may use these data to develop novel insights into mechanisms of discrimination reduction by identifying the "key ingredients" of effective behavior change. In short, a research contest design will be highly generative for finding what strategies best reduce discriminatory behavior and can facilitate future investigations into the psychological processes that give rise to their effectiveness.

Contest Overview

Our contest had the goal of comparing a variety of interventions that researchers believed would reduce discrimination in the Judgment Bias Task (JBT; Axt et al., 2018). The JBT is a decision-making task that has been shown to consistently reveal discrimination in decision-making (Axt et al., 2018, 2019; Axt & Johnson, 2021; Axt & Lai, 2019; Axt et al., 2021). During the task, participants are presented with applications for a hypothetical academic honor society that contain both relevant information (i.e., qualifications) and irrelevant information (i.e., a picture of the applicant). More specifically, the JBT used in the present work contained four relevant pieces of information for evaluating applicants – GPA in science (1-4), GPA in humanities (1-4), interview score (1-100), and recommendation letter score (bad, fair, good, outstanding) – and one irrelevant piece of information: the applicant's photo. Even though the difference between more qualified and less qualified applicants is small, the applications are

designed in such a way that some are objectively better than others. Particularly, the standardized sums of all four qualification indicators were the same across all "less qualified" candidates and similarly across all "more qualified" candidates, which allows for responses to be compared to an objectively correct answer (Axt et al., 2018).

Previous work using the JBT has shown that the magnitude of discrimination can be viewed as the product of two processes using Signal Detection Theory (SDT; Green & Swets, 1966). First, participants may make errors in judgment, which leads to noise (Axt & Lai, 2019). That is, participants may poorly differentiate between qualified and unqualified candidates, resulting in more incorrect decisions. In SDT, this ability to differentiate between more and less qualified applicants is called sensitivity (d'); greater sensitivity means less noise. Given the 50/50 split between more and less qualified applicants in the JBT used here, sensitivity is highly correlated with overall accuracy (r = .97 across all participants).

Second, participants may have a distribution of errors that reveals social favoritism (Axt & Lai, 2019). For example, errors may show that participants are too lenient towards physically attractive applicants (i.e., more errors of falsely accepting less qualified people) and are too stringent towards less physically attractive applicants (i.e., more errors of falsely rejecting more qualified people). In this context, an uneven distribution of the kind of errors committed is called bias, and SDT quantifies this effect by comparing the response criterion value for more versus less physically attractive applicants. Criterion refers to decision threshold (i.e., what level of qualification is needed to elicit an "accept" decision). Lower values of criterion reflect more leniency in the selection process, indicating that applicants do not need to be as qualified in order to receive an "accept" decision. Prior JBT studies consistently show lower criterion values for more relative to less physically attractive applicants (Axt et al., 2018). We build off these

prior studies by also using a JBT that focuses on discrimination concerning physical attractiveness.

The magnitude of discrimination is then a combination of how much noise and bias there is in judgment; whereas noise concerns the total number of errors made, bias is a proportional measure that indicates the rate at which the errors made favor one group over another. Prior work (Axt & Lai, 2019) has found that bias and noise are differently impacted by certain interventions. When unfair treatment is present in judgment, interventions that reduce decision bias (i.e., the relative degree of preference for a particular group) and interventions that reduce decision noise (i.e., the total number of errors/unfair judgments made) are both effective routes towards reducing overall discrimination. In other words, successful interventions may reduce the degree to which one group is favored over another in judgment either by leaving overall amount of errors made (i.e., noise) unchanged but reducing the proportional spread of errors in favor of a certain group (i.e., bias) or by leaving the proportional spread of unequitable errors unchanged but reducing the overall amount of judgment errors committed. In the online supplement, we document the results of a simulated, 64-trial JBT that illustrates how three different hypothetical interventions – one that only increases sensitivity, one that only reduces criterion bias, and one that does both – can be equally effective at reducing discrimination (here, lessening the degree to which acceptance decisions are given towards more attractive applicants relative to equally-qualified less attractive applicants). In all, interventions in our research contest were considered effective if they reduced either noise or bias on the JBT, though we considered interventions as ideal and most effective if they simultaneously reduced bias and increased sensitivity.

To solicit interventions for the contest, we advertised on various platforms, such as social media, discussion boards, listservs, and newsletters for relevant professional societies. In total, we received 30 interventions, which were submitted by researchers from around the world. Interventions varied greatly in terms of the psychological mechanism employed and methodological demands. However, each intervention had to respect the contest's criteria, specifically that 1) interventions could not explicitly mention physical attractiveness, 2) the total duration of the intervention needed to be less than seven minutes, and 3) the intervention could not involve changes to the JBT itself (e.g., requiring a 10-second delay before responses could be made).

These criteria were put in place for several reasons. First, interventions could not mention physical attractiveness, as a goal of the contest was to develop generalizable strategies that could be applied to various forms of discrimination, including potentially reducing discrimination along multiple dimensions simultaneously. Prior work using the JBT (Axt et al., 2019) has found that merely raising awareness about one form of discrimination (e.g., based on physical attractiveness) has a narrow effect on behavior, with no impact on other forms of discrimination that may be operating simultaneously (e.g., based on political affiliation). Second, interventions needed to be less than seven minutes to enhance comparability across interventions and minimize dropout among our online participant samples. Third, interventions could not change the JBT because doing so would impair comparability across intervention (i.e., interventions would no longer have the exact same outcome measure).

In total, we tested submitted interventions across four rounds of data collection and more than 20,000 participants. The first two rounds of the contest (Studies 1-2) focus on intervention effectiveness for a single form of discrimination (i.e., based on physical attractiveness) using

convenience samples. Study 3 builds on the findings from the prior rounds and explores how interventions previously identified as reducing noise and/or bias would fare in a context where two potential sources of discrimination (i.e., based on physical attractiveness and political affiliation) exist simultaneously. Lastly, the fourth round (Study 4) investigates the generalizability of the findings across a sample of participants with self-reported hiring experience.

Studies 1-2

Methods

Participants

Study 1 participants came from Project Implicit (implicit.harvard.edu), a non-profit organization and online research laboratory. A total of 12,519 participants completed at least all trials in the JBT (i.e., provided full data on our main outcome variable). As in the task validation studies (e.g., Axt et al., 2018), data from participants who did not fully complete the JBT's 64 trials or who had an acceptance rate of less than 20% or more than 80% were removed from the analysis, in addition to participants who either accepted or rejected all of the more or less physically attractive applicants, respectively (Axt et al., 2018)[1]. The resulting sample was N = 11,196 (M_{age} = 38.4, SD_{age} = 15.2, 58.2% White, 65.7% female). Data collection continued until there was an average of 350 eligible participants in every condition, which provided more than 80% power to detect an effect as small as d = .21 when comparing each intervention to the control condition. See https://osf.io/m2a9w/ for pre-registration of all methods, measures, data

[1] When including all participants with full JBT data, conclusions from only 3 out of the 99 (3.03%) reported analyses changed across all four studies.

exclusion practices, and analyses. See https://osf.io/wk2s9/ for materials, data, and analysis syntax for all studies.

Following this first round of data collection, Study 2 was a direct replication using participants from Prolific (https://www.prolific.co), allowing for a test of whether Study 1 results (which used volunteer participants) would replicate among a sample of paid participants. Aside from investigating whether the effects of interventions identified in Study 1 as successfully reducing decision noise, decision bias, or both outcomes would replicate in a new sample, Study 2 limited the risks of Type I errors. That is, even if no interventions actually changed JBT decision-making, testing 30 interventions across two outcomes each at $\alpha = .05$ could lead to three Study 1 interventions being falsely identified as effective at changing at least one outcome based on chance.

For Study 2, we included 11 interventions from Study 1. Nine interventions had either reliably reduced biases in response criterion, reliably increased sensitivity, or impacted both outcomes simultaneously. One additional intervention had a marginally significant effect on increasing sensitivity in Study 1, and one had an abnormally low sample size in Study 1. A total of 4,731 participants completed Study 2. We excluded participants based on the same criteria as Study 1, resulting in a final sample of N = 4,446 (M_{age} = 37.7, SD_{age} = 13.4, 78.6% White, 50.8% female).[2] We targeted an average of at least 400 participants per condition (final average N = 404 per condition), which provided more than 80% power to detect an effect as small as d = .20 when comparing each intervention to the control condition.

[2] The direct replication of one intervention (Mindfulness Exercise) was conducted separately due to an analysis error in Study 1. For this replication, participants were randomly assigned to a control condition or the Mindfulness Exercise intervention, using Prolific. The total sample was N = 842. After applying the same data-cleaning procedures as the other studies, the final sample was N = 811 (M_{age} = 39.2, SD_{age} = 14.0, 79.2% White, 47.4% female).

Procedure

Study 1 participants reported demographics when first registering for the Project Implicit research pool. Participants in Study 1 were then assigned to one of 31 conditions (one control condition that completed a JBT following standard task instructions and 30 intervention conditions), and participants in Study 2 were assigned to one of 12 conditions (one control condition plus 11 intervention conditions). After receiving their intervention, participants in both studies completed the JBT and then a short self-report questionnaire. Study 2 participants then completed a seven-item demographics questionnaire as well as an attention check item. All participants were debriefed and received feedback on their JBT performance.

Judgment Bias Task. The JBT employed in these studies involved evaluating 64 applicants (Axt et al., 2018) for an academic honor society. In particular, the study used a JBT investigating discrimination based on physical attractiveness, which past research has shown to be a significant source of bias (Feingold, 1992).

Each application vignette was made up of a picture of the applicant, as well as four qualification indicators: GPA in science (1-4), GPA in humanities (1-4), interview score (1-100), and recommendation letter score (bad, fair, good, excellent). Within the JBT, applicants were then evaluated in a two (more physically attractive, less physically attractive) by two (more qualified, less qualified) within-subjects design. To manipulate physical attractiveness, we used the same pictures from prior JBT studies, which had been rated for differences in perceived attractiveness (Axt et al., 2018). These pictures were all of White, college-aged people who were smiling. Each level of physical attractiveness (i.e., high vs. low) had an equal number of men versus women.

To manipulate applicant qualifications, we first placed all qualifications on the same 1-4 scale. GPAs were already on a 1-4 scale, interview scores were divided by 25, and recommendation letters were scored such that *Bad* = 1, *Fair* = 2, *Good* = 3, and *Excellent* = 4. We then built unique qualification combinations such that the sum of the qualifications for a given applicant would add up to either 14 (more qualified) or 13 (less qualified). For example, a "more qualified" candidate could have a 3.8 science GPA, a 3.3 humanities GPA, "good" recommendation letters (value of 3), and a 97.5 interview score (value of 3.9 when divided by 25). Another equally "more qualified" candidate could, alternatively, have a 3.1 science GPA, a 3.4 humanities GPA, "excellent" recommendation letters (a value of 4), and an 87.5 interview score (value of 3.5 when divided by 25). Participants were randomly assigned to one of 12 JBT orders; across orders, each application was equally likely to be paired with a more versus less physically attractive face.

The JBT had two phases: an encoding task that passively displayed each application for one second, and a test phase in which participants chose to accept or reject each applicant (with a 15-second timeout per trial). After the encoding phase, participants were instructed to accept approximately half of the applicants, selecting those whom they deemed to be "more qualified" over others. Interventions were allowed to vary in terms of administering components of the intervention before versus after the encoding phase; all interventions preceded the test phase.

Self-report questionnaire. Participants completed three self-report items that have been used in previous JBT studies (Axt et al., 2018, 2019; Axt & Lai, 2019). First, participants were asked about their perceived and desired performance on the task. They also reported their explicit attractiveness attitudes. These items were not included in any confirmatory analyses in our pre-registration but were added to data collection to facilitate potential exploratory analyses

in future work (e.g., whether interventions that reduced discrimination were mediated by changes to desired JBT performance).

Calculating Intervention Effectiveness

For ease of interpretation, we determined intervention effectiveness as a combination of how well the manipulation either reduced attractiveness-based biases in response criterion or increased overall sensitivity. Specifically, interventions were judged as more effective when they reliably increased sensitivity or reliably reduced attractiveness-based differences in criterion at $p < .05$. An intervention's total effectiveness score was then the sum of the Cohen's *d* effect size for the degree to which the intervention reliably impacted sensitivity and/or criterion bias (see online supplement for specific instructions given to researchers about how interventions would be evaluated). For interventions that did not reliably change either outcome, we averaged the Cohen's *d* effect sizes for reducing criterion bias and increasing sensitivity, though these interventions were always ranked after any intervention that reliably changed either outcome. Using Study 1 results, this method was used to determine authorship order (excluding the first and two last authors).

Background and Results

Table 1 lists the sample size, acceptance rate, and accuracy for each condition in Study 1, in addition to descriptive statistics for sensitivity, criterion for less attractive applicants, criterion for more attractive applicants, and a criterion bias difference score (calculated by subtracting the criterion for more physically attractive applicants from the criterion for less physically attractive applicants, such that higher values mean more leniency towards more less attractive applicants). Finally, Table 1 also lists the results of a within-subjects *t*-test for each condition comparing the criterion for more versus less physically attractive applicants. See

Table 2 for the same information regarding Study 2. In both studies, control conditions showed accuracy = 67.8%. In addition, control conditions in both studies showed a robust criterion bias (Study 1 control condition bias d = .24, Study 2 control condition bias d = .29).

In line with our pre-registration, our primary analyses compared each intervention to the control condition on overall sensitivity as well as the criterion bias difference score.[3] See Figure 1 (Study 1) and Figure 2 (Study 2) for graphical displays of results concerning the Cohen's *d* effect size of each intervention's effect on sensitivity and the criterion bias difference score relative to the control condition. To facilitate presentation of methods and results, all interventions were grouped into seven categories, which were agreed on post-hoc by the first and two last authors. Specifically, interventions fell into the following categories: 1) accountability, 2) association training, 3) bias awareness, 4) counter-stereotypical exemplar, 5) general self-reflection, 6) evaluation criteria (provided), and 7) evaluation criteria (self-determined). See Table 3 for more information about each category, as well as the interventions included in each category. Given the number of interventions, in-text summaries and rationales for each intervention are condensed, and researchers were given the opportunity to present a longer rationale in the online supplement. The online supplement also contains the full materials for each intervention. Finally, readers can test the JBT at this link: https://tinyurl.com/3t5vnyv9, and review each intervention at this link: https://tinyurl.com/3nnm3j6n.

[3] Whenever Levene's Tests showed reliable differences in variances, we report results with equal variances not assumed (conclusions never change when assuming equal variances).

Table 1

Descriptive and Test Statistics for Each Condition in Study 1

Condition	*N*	Acceptance Rate	Accuracy	Sensitivity	Criterion Bias Difference	More Attractive Criterion	Less Attractive Criterion	Criterion Comparison
Control	400	51.3%	66.5%	0.95 (.97)	0.10 (.43)	-0.09 (.45)	0.02 (.46)	$t(399) = 4.78$, $d = 0.24$, $p < .001$
Justification Instructions	406	52.2%	66.6%	0.97 (.57)	0.09 (.46)	-0.11 (.46)	-0.02 (.48)	$t(405) = 4.05$, $d = 0.20$, $p < .001$
Attractive = Harmful Movie	365	50.9%	66.1%	0.92 (.55)	0.09 (.46)	-0.07 (.43)	0.02 (.46)	$t(364) = 3.52$, $d = 0.18$, $p < .001$
Blind Evaluation Instructions	439	51.6%	67.4%	1.00 (.56)	0.06 (.41)	-0.07 (.44)	-0.02 (.43)	$t(438) = 2.89$, d = 0.14, $p < .01$
Criteria from Min-Max Values	301	52.0%	67.3%	1.02 (.60)	0.04 (.37)	-0.08 (.44)	-0.04 (.45)	$t(300) = 1.85$, $d = 0.11$, $p = .066$
Criteria from Mean Values	295	48.9%	65.7%	0.94 (.56)	0.04 (.40)	0.03 (.49)	0.06 (.52)	$t(294) = 1.56$, $d = 0.09$, $p = .120$
Jury Instructions	445	52.0%	67.2%	1.00 (.54)	0.07 (.39)	-0.09 (.44)	-0.02 (.48)	$t(444) = 3.88$, $d = 0.18$, $p < .001$
Orchestra Case	440	52.0%	67.0%	0.98 (.57)	0.10 (.48)	-0.11 (.45)	-0.01 (.46)	$t(439) = 4.49$, $d = 0.21$, $p < .001$
Bias Blind Spot	389	52.8%	67.0%	0.98 (.52)	0.06 (.44)	-0.11 (.47)	-0.05 (.44)	$t(388) = 2.80$, $d = 0.14$, $p < .01$
Imagined Contact	240	51.4%	65.5%	0.89 (.58)	0.15 (.50)	-0.11 (.47)	0.03 (.47)	$t(239) = 4.57$, $d = 0.30$, $p < .001$
Personal Information Avoidance Rule	417	51.2%	66.2%	0.94 (.50)	0.09 (.38)	-0.08 (.47)	0.01 (.45)	$t(416) = 4.90$, $d = 0.24$, $p < .001$
Norm Information	474	50.9%	66.5%	0.96 (.61)	0.12 (.44)	-0.08 (45)	0.04 (.46)	$t(473) = 6.02$, $d = 0.28$, $p < .001$
Personalized Moral Concern	437	52.4%	67.2%	1.00 (.55)	0.07 (.40)	-0.10 (.44)	-0.04 (.47)	$t(436) = 3.47$, $d = 0.17$, $p < .001$
Minimal Threshold Criteria	325	47.9%	65.4%	0.92 (.52)	0.07 (.40)	0.04 (.51)	0.10 (.50)	$t(324) = 2.96$, $d = 0.16$, $p < .01$
Automatic Mental Processes	473	52.3%	67.2%	0.99 (.52)	0.06 (.46)	-0.10 (.45)	-0.04 (.46)	$t(472) = 2.91$, $d = 0.13$, $p < .01$

Associative Learning Paradigm	161	51.7%	65.4%	0.89 (.56)	0.05 (.48)	-0.07 (0.50)	-0.02 (.50)	$t(160) = 1.28$, $d = 0.10$, $p = .204$
Moral First Impressions	413	53.7%	66.5%	0.96 (.52)	0.06 (.44)	-0.14 (.47)	-0.09 (.48)	$t(412) = 2.57$, $d = 0.13$, $p < .05$
Vivid Narrative Exercise	421	51.2%	67.9%	1.03 (.55)	0.09 (.41)	-0.08 (.43)	0.01 (.45)	$t(420) = 4.53$, $d = 0.22$, $p < .001$
Like-Dislike Writing	320	52.1%	65.6%	0.90 (.58)	0.08 (.52)	-0.11 (.48)	-0.02 (.48)	$t(319) = 2.91$, $d = 0.16$, $p < .01$
Cultural Self-Awareness	272	51.8%	66.8%	0.95 (.55)	0.07 (.40)	-0.09 (.40)	-0.02 (.41)	$t(271) = 2.83$, $d = 0.17$, $p < .01$
Two-Out-of-Three Rule	441	47.3%	79.5%	1.83 (.77)	0.04 (.29)	0.10 (.34)	0.14 (.33)	$t(440) = 3.07$, $d = 0.16$, $p < .01$
WOOP-inspired Video	252	50.3%	68.5%	1.10 (.57)	0.08 (.40)	-0.04 (.46)	0.04 (.46)	$t(251) = 3.05$, $d = 0.20$, $p < .01$
Mindfulness Exercise	237	52.3%	67.9%	1.06 (.61)	0.07 (.42)	-0.10 (.47)	-0.03 (.45)	$t(236) = 2.49$, $d = 0.16$, $p < .05$
Trial-and-Error Feedback	344	52.5%	67.9%	1.03 (.58)	0.07 (.40)	-0.11 (.42)	-0.04 (.42)	$t(343) = 3.16$, $d = 0.17$, $p < .01$
Separate Judgment Exercise	404	52.0%	67.2%	1.00 (.54)	0.12 (.46)	-0.11 (.47)	0.00 (.47)	$t(403) = 5.03$, $d = 0.25$, $p < .001$
Qualified Brief IAT	299	51.6%	66.6%	0.95 (.50)	0.09 (.37)	-0.09 (.45)	0.00 (.43)	$t(298) = 4.03$, $d = 0.23$, $p < .001$
Criteria Reinforcement Exercise	334	39.3%	74.6%	1.67 (.70)	0.03 (.29)	0.36 (.50)	0.39 (.49)	$t(333) = 2.11$, $d = 0.12$, $p < .05$
Single-Criterion Exercise	300	53.7%	66.2%	0.96 (.55)	0.03 (.37)	-0.13 (.49)	-0.1 (.47)	$t(299) = 1.25$, $d = 0.07$, $p = .211$
Similarity-Attraction Paradigm	411	51.5%	66.6%	0.95 (.55)	0.10 (.46)	-0.09 (.47)	0.01 (.46)	$t(410) = 4.54$, $d = 0.22$, $p < .001$
Propositional-Statistical Learning	325	51.8%	65.3%	0.84 (.36)	0.04 (.27)	-0.07 (.35)	-0.03 (.34)	$t(324) = 2.55$, $d = 0.14$, $p < .05$
Morality-Competence Exercise	409	51.9%	66.9%	0.98 (.56)	0.08 (.43)	-0.09 (.45)	-0.01 (.45)	$t(408) = 2.52$, $d = 0.12$, $p < .05$

Table 2

Descriptive and Test Statistics for Each Condition in Study 2

Condition	*N*	Acceptance Rate	Accuracy	Sensitivity	Criterion Bias Difference	More Attractive *c*	Less Attractive *c*	Criterion Comparison
Control	417	51.3%	67.8%	1.02 (.54)	0.12 (0.41)	-0.10 (.42)	0.02 (.41)	$t(416) = 5.99$, $d = 0.29$, $p < .001$
Criteria from Min-Max Values	389	51.4%	66.7%	0.96 (.58)	0.10 (0.39)	-0.09 (.39)	0.02 (.41)	$t(388) = 5.35$, $d = 0.27$, $p < 001$
Criteria from Mean Values	376	48.1%	67.4%	1.04 (.63)	0.04 (0.34)	0.05 (.39)	0.10 (.43)	$t(375) = 2.51$, $d = 0.13$, $p = .013$
Associative Learning Paradigm	405	51.7%	66.0%	0.90 (.52)	0.12 (0.43)	-0.11 (.32)	0.01 (.32)	$t(404) = 5.51$, $d = 0.27$, $p < .001$
Vivid Narrative Exercise	412	51.8%	67.5%	0.99 (.48)	0.10 (0.42)	-0.10 (.43)	0.00 (.43)	$t(411) = 5.06$, $d = 0.25$, $p < .001$
Two-Out-of-Three Rule	445	47.5%	80.3%	1.87 (.76)	0.02 (0.27)	0.10 (.41)	0.13 (.42)	$t(444) = 1.77$, $d = 0.08$, $p = .078$
WOOP-inspired Video	377	52.3%	68.3%	1.06 (.56)	0.10 (0.37)	-0.12 (.48)	-0.02 (.49)	$t(376) = 5.04$, $d = 0.26$, $p < .001$
Mindfulness Exercise	380	52.3%	66.9%	0.97 (.51)	0.11 (.45)	-0.12 (.42)	-0.01 (.47)	$t(379) = 4.92$, $d = 0.25$, $p < .001$
Trial-and-Error Feedback	440	51.1%	69.2%	1.11 (.58)	0.08 (.37)	-0.08 (.39)	0.01 (.42)	$t(439) = 4.82$, $d = 0.23$, $p < .001$
Criteria Reinforcement Exercise	415	39.2%	73.7%	1.62 (.76)	0.01 (0.30)	0.38 (.30)	0.39 (.31)	$t(414) = 0.69$, $d = 0.03$, $p = .492$
Single-Criterion Exercise	375	51.1%	65.0%	0.85 (.53)	0.06 (0.38)	-0.06 (.42)	-0.01 (.42)	$t(374) = 2.89$, $d = 0.15$, $p = .004$
Propositional-Statistical Learning	395	52.0%	64.5%	0.79 (.38)	0.004 (0.27)	-0.06 (.42)	-0.05 (.42)	$t(394) = 0.28$, $d = 0.01$, $p = .783$

Figure 1. Difference in Criterion Bias and Sensitivity with the Control Condition for each Intervention in Study 1.

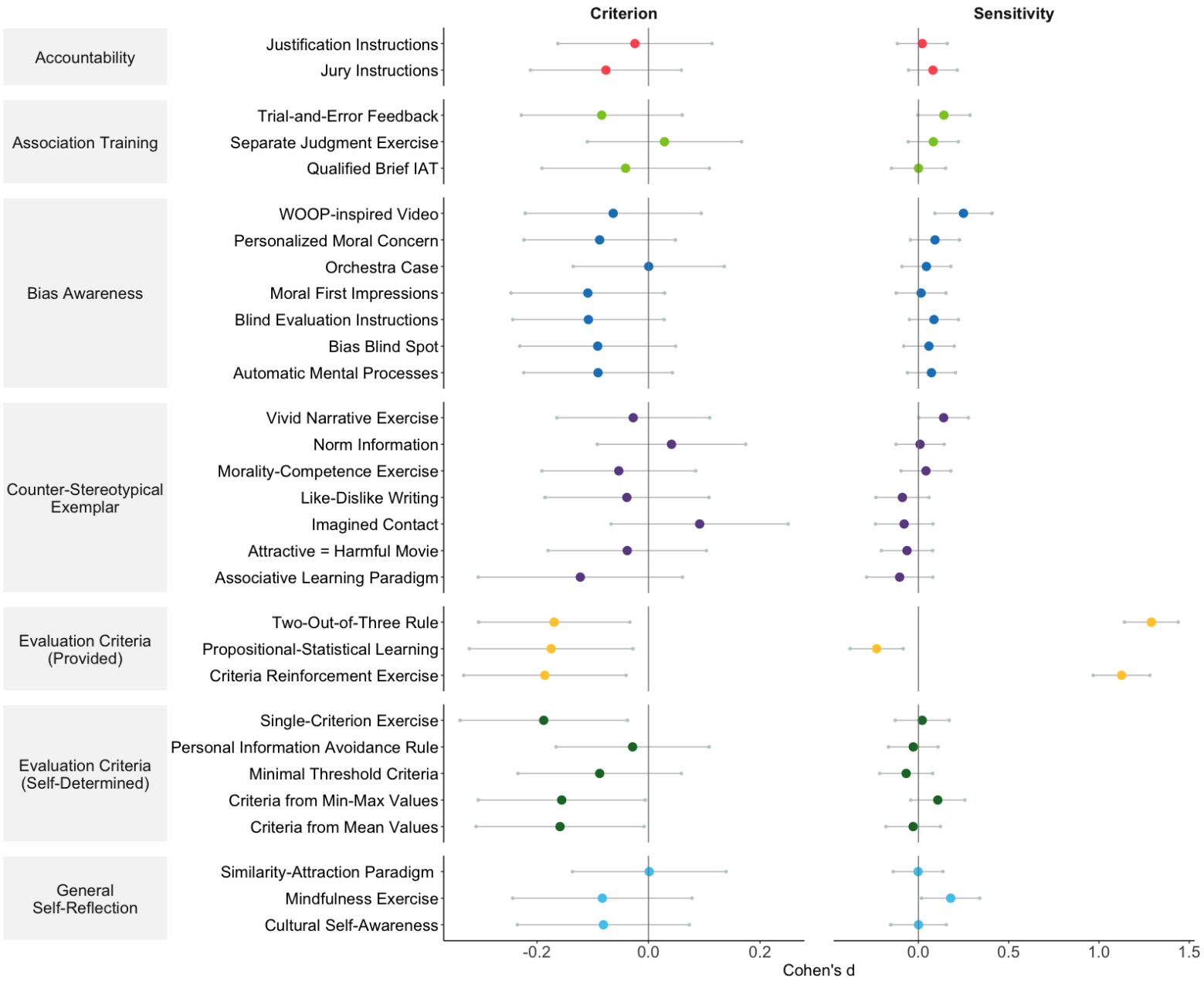


Figure 2. Difference in Criterion Bias and Sensitivity with the Control Condition for each Intervention in Study 2.

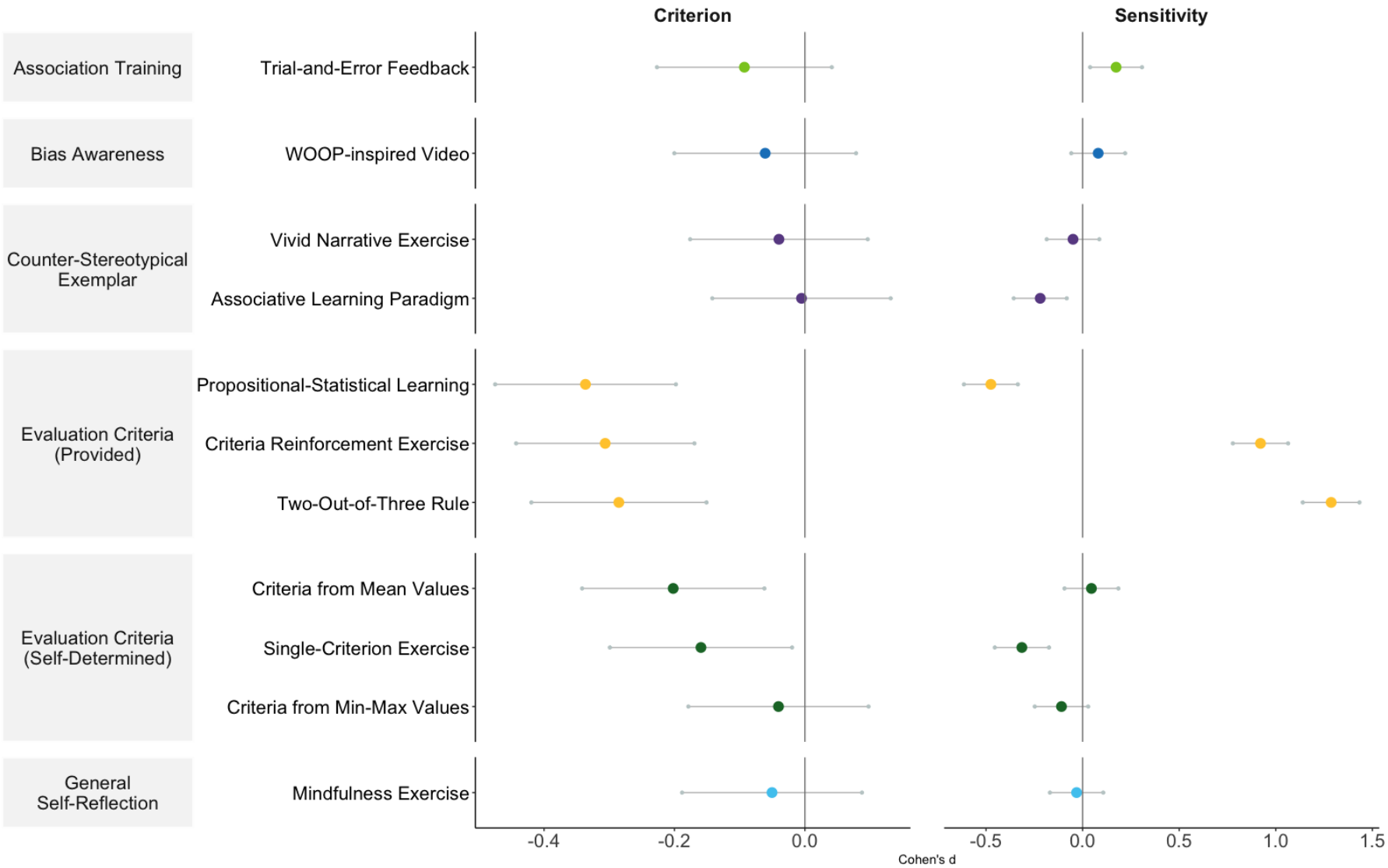


Table 3.
List of Intervention Categories, Descriptions, and Names.

Category	Description	Intervention Name
Accountability	This category included all interventions that fostered a sense of accountability on the participant's part with regard to the choices they would be making in the JBT.	Justification Instructions Jury Instructions
Association Training	This category included all interventions that reinforced an association between concepts or groups that would reduce bias in judgement.	Trial-and-Error Feedback Separate Judgment Exercise Qualified Brief IAT
Bias Awareness	This category included all interventions that either provided information about how judgments can be biased or gave general strategies to overcome biases and make more objective judgments.	WOOP-inspired Video Personalized Moral Concern Orchestra Case Moral First Impressions Blind Evaluation Instructions Bias Blind Spot Automatic Mental Processes
Counter-Stereotypical Exemplar	This category included all interventions that showed or described situations in which people acted in opposition to a stereotype typically held towards that group. In the context of this study, these would include more physically attractive people engaging in negative behaviors and/or less physically attractive people engaging in positive behaviors.	Vivid Narrative Exercise Norm Information Morality-Competence Exercise Like-Dislike Writing Imagined Contact Attractive = Harmful Movie Associative Learning Paradigm
Evaluation Criteria (Provided)	This category included all interventions that gave participants specific evaluation criteria or a decision-making rule to use when completing the JBT.	Two-Out-of-Three Rule Propositional-Statistical Learning Criteria Reinforcement Exercise
Evaluation Criteria (Self-Determined)	This category included all interventions that required the participant to form their own evaluation criteria or decision-making rule when completing the JBT.	Single-Criterion Exercise Personal Information Avoidance Rule Minimal Threshold Criteria Criteria from Min-Max Values Criteria from Mean Values
General Self-Reflection	This category included all interventions that invited participants to reflect on themselves and their thoughts for some time before moving on to the JBT.	Similarity-Attraction Paradigm Mindfulness Exercise Cultural Self-Awareness

Accountability Interventions

Jury Instructions (Jacqueline M. Chen and Teneille Brown). Justice requires jurors to base their verdicts on legally relevant characteristics. To achieve this goal, judges provide concise legal instructions to juries (Jones et al., 2022). These instructions are thus employed as de-biasing strategies – to reduce the impact of prejudicial information. Most studies have found that instructions for jurors to disregard information backfire—increasing the impact of the to-be-disregarded information (Wissler et al., 2000). Even so, some remain hopeful that specialized instructions have mitigated implicit racial bias in a subset of cases (Capers et al., 2018). To investigate this idea, the 'Jury Instructions' intervention asked participants to review a common jury instruction to test whether it effectively mitigated the impact of physical appearance on judgment. This intervention did not increase overall sensitivity, $t(843) = -1.17$, $p = .244$, $d = .08$ and did not reduce criterion biases, $t(843) = -1.11$, $p = .268$, $d = -.08$.

Justification Instructions (Jeremy D. Gretton). Work from Lerner and Tetlock (2003) shows that accountability comes in many forms. In their review, Lerner and Tetlock (2003) note that accountability can reduce bias that is due to insufficient "self-critical attention to the judgment process," to the extent that debiasing does not require training regarding how to make decisions. This intervention instructed participants: "Please note that following this task you will be asked to justify your decisions. Specifically, you will be shown one of the applicants. Then, you will be asked to reflect on how you evaluated them and what criteria you used to either accept or reject this applicant." Thus, after having completed the JBT, participants were asked to write down why they accepted or rejected one applicant, selected at random. This intervention did not increase overall sensitivity, $t(804) = -0.3$, $p = .761$, $d = .02$ and did not reduce criterion biases, $t(804) = -0.35$, $p = .729$, $d = -.02$.

Association Training Interventions

Trial-and-Error Feedback (Jessica Sim and Nazia Khan). Feedback can have a positive impact on performance (Kluger & DeNisi, 1996). When evaluating job performance, practice and feedback can reduce rater errors (e.g., halo effect) and improve rater accuracy (Smith, 1986). In this intervention, participants completed a 32-trial practice JBT that used novel faces and applications. Participants received feedback when submitting a response and were reminded to use all four qualifications equally after an incorrect choice. In Study 1, this intervention fell just short of reliably increasing overall sensitivity, $t(742) = -1.92$, $p = .055$, $d = .14$, and did not reduce criterion biases, $t(742) = 1.14$, $p = .255$, $d = -.08$. Given these ambiguous results, we included the intervention in Study 2, where it did increase overall sensitivity, $t(855) = -2.53$, $p = .012$, $d = .17$ and did not reduce criterion bias, $t(855) = -1.36$, $p = .176$, $d = -.09$.

Separate Judgment Exercise (Aishwarya Kumar and Ranjavati Banerji). Advertising strategies tend to exploit harmful physical and gender stereotypes to sell products to target consumer groups. This intervention was interested in testing the extent to which young consumers are more conscious of potential biases and resistant to social discrimination relating to physical appearance and gender identity in their consumer habits. In a task similar to the JBT, participants selected the most suitable actors to advertise a pair of sneakers. The pictures were of novel male and female faces that had previously been rated as "more physically attractive" or "less physically attractive" in pre-testing. The qualification information included characteristics like audition performance, professional experience, education, etc. Participants first viewed each applicant before starting the 16-trial selection process. This intervention did not increase overall sensitivity, $t(802) = -1.17$, $p = .242$, $d = .08$, and did not reduce criterion bias, $t(802) = -0.40$, $p = .686$, $d = .03$.

Qualified Brief IAT (Bethany Teachman). The goal of this intervention was to promote strong associations between the characteristics of a qualified candidate (e.g., high GPA) and that candidate being selected. The intervention used a modified Brief IAT (BIAT; Sriram & Greenwald, 2009) with "Qualified Candidate" as the target category and "Select" (vs "Reject") as the attribute category labels. Unlike a traditional BIAT, in which the category pairings switch across blocks, all four blocks in this task (80 total trials) paired "Qualified Candidate" with "Select." Stimuli were chosen so that the Qualified Candidate category was composed of the same indicators of qualified applicants on the JBT (Science GPA > 3.8, Humanities GPA > 3.0, Good Recommendation, Interview Score > 80). The unlabeled, contrasting background stimuli were *Incompetent, Unqualified, Bad grades, Ineffective,* and the attribute category stimuli were synonyms for Select (*Select, Say Yes, Accept, Choose*) and Reject (*Skip, Say No, Ignore, Reject*). This intervention did not increase overall sensitivity, $t(697) = -0.01$, $p = .995$, $d < .01$, and did not reduce criterion biases, $t(697) = -0.54$, $p = .592$, $d = -.04$.

Bias Awareness Interventions

WOOP-inspired Video (Brian O'Shea and Michael Bang Petersen). This intervention consisted of a whiteboard video animation (5.5 minutes) that included a breathing/relaxation segment, followed by a WOOP-inspired (Wish, Outcome, Obstacle, Plan; Oettingen, 2015) active participation segment. First, participants learned that "Most people express a desire to treat others fair and equally" and reviewed why people may struggle to meet these ideals (e.g., unconscious biases). Participants then twice repeated the following statement: "I wish to treat the job candidates fairly and only base selection on their four scores." Next, participants imagined the best outcomes if their wish came true. Finally, participants developed implementation intention strategies to overcome any obstacles (Bieleke et al., 2021). The

instructions encouraged participants to "complete the task the same way a scientist or mathematician would, using rationality, objectivity, and analytic skills," or to leverage your arousal from the stress by "focussing on the calculation strategy to reduce selection bias." After the video, participants had an open text box to write out the plan they would adopt for the upcoming selection task. This intervention did increase overall sensitivity, $t(650) = -3.10$, $p = .002$, $d = .25$, and did not reduce criterion biases, $t(650) = 0.79$, $p = .432$, $d = -.06$. Since the intervention increased sensitivity in Study 1, we included it in Study 2. In Study 2, this intervention did not increase overall sensitivity, $t(792) = -1.13$, $p = .257$, $d = .08$ and did not reduce criterion biases, $t(792) = -0.86$, $p = .393$, $d = -.06$.

Bias Blind Spot (Haewon Yoon, Carey Morewedge, and Irene Scopelliti). People perceive themselves to be less biased than their peers, and these self-assessments do not relate to their actual susceptibility to bias or decision-making abilities (i.e., bias blind spot; Pronin et al., 2002; Scopelliti et al., 2015). This bias blind spot has detrimental consequences for judgment and decisions. This intervention provided information about the bias blind spot, explaining how it is easier to detect someone else's biases than our own (Pronin & Kugler, 2007). The intervention encouraged participants to recognize their own bias by examining their judgment outcomes as if they were external observers. Following the intervention, participants answered two comprehension checks on why it is harder to detect one's own bias compared to others and how the bias blind spot effect can be reduced. This intervention did not increase overall sensitivity, $t(787) = -0.82$, $p = .415$, $d = .06$, and did not reduce criterion biases, $t(787) = -1.28$, $p = .202$, $d = -.09$.

Personalized Moral Concern (Luiza Almeida Santos and Jan G. Voelkel). This intervention makes a moral argument in favor of unbiased selection. First, we measured

participants' political ideology. Then, we presented the participants with an argument that appeals to the moral values that are typically strongly endorsed by the ideological group with which the participants identified (Feinberg & Willer, 2019). Research suggests that the most important moral principles for liberals are minimizing harm and ensuring fairness, whereas conservatives also moralize in-group loyalty, respect for authority, and sanctity (Graham et al., 2009). Building on this work, this intervention asked participants to read an essay emphasizing that unbiased selection is consistent with their moral concerns. To assign participants to text that aligned with their moral values, the intervention first asked participants to report political orientation on a scale from 1 (*extremely liberal*) to 7 (*extremely conservative*), with 4 being the neutral point. For liberal participants (who scored between 1 and 3), unbiased selection was framed as a means of ensuring fairness, while for conservative participants (who scored between 5 and 7), unbiased selection was framed as an expression of patriotism and commitment to American ideals. Because research has found that conservative values resonate with moderate participants (Voelkel et al., 2021), neutral participants were assigned the conservative value essay. This intervention did not increase overall sensitivity, $t(835) = -1.33$, $p = .186$, $d = .09$, and did not reduce criterion biases, $t(835) = -1.26$, $p = .206$, $d = -.09$.[4]

Orchestra Case (Landon Schnabel). Evaluators often try to make rational decisions and pick the best candidates, focusing only on relevant "objective" criteria while trying to ignore

[4] Given that some components of the intervention relied on invoking American ideals, we ran a follow-up, exploratory analysis keeping only American citizen control condition and intervention participants (total $N = 597$). This sample showed no increase in overall sensitivity, $t(595) = -0.86$, $p = .392$, $d = .07$, and no change in criterion biases, $t(595) = -0.65$, $p = .515$, $d = -.05$. We also divided the American-only sample into Conservative and Liberal subgroups. The American-Conservative subgroup showed no increase in overall sensitivity, $t(113) = -0.57$, $p = .570$, $d = -.11$, and did not reduce criterion biases, $t(113) = -1.38$, $p = .171$, $d = -.26$. The American-Liberal subgroup also showed no increase in overall sensitivity, $t(316) = -1.74$, $p = .08$, $d = .20$, and did not reduce criterion biases, $t(316) = 0.22$, $p = .823$, $d = .03$.

biasing information. However, much of cognition is automatic and conscious attempts to "ignore" irrelevant information occur too late in the thought process (Miles et al., 2019; Rivera, 2015). Regardless of intentions, evaluators develop quick first impressions—frequently based on stereotypes about immaterial characteristics—and end up discriminating based on social factors such as gender and race (Pager, 2003; Quadlin, 2018). This intervention had participants read a short article about subtle discrimination and how orchestras sought to address it by putting up screens and otherwise blinding evaluators to irrelevant information (Goldin & Rouse, 2000). It then invited respondents to implement a similar approach, using their hand as a "screen" to block irrelevant information. This intervention did not increase overall sensitivity, $t(838) = -0.64$, $p = .522$, $d = .04$ and did not reduce criterion biases, $t(838) = 0.01$, $p = .995$, $d < .01$.

Automatic Mental Processes (Maddalena Marini). Studies have shown that raising awareness of bias and asking for more deliberative judgments reduces discrimination (Axt & Johnson, 2021; Axt & Lai, 2019; Pope et al., 2018), but only when the social category was named explicitly (Axt et al., 2019). In this intervention, participants read an article raising awareness of automatic mental processes that can guide our behavior and influence how we form impressions of others. Participants were also presented with a picture representing two groups of people, one composed of more physically attractive members and the other with less physically attractive members. Then, they were informed that, even if all these people have equally strong qualifications, studies showed that the one group (pictured as several more attractive individuals) was typically judged as more competent than the other group (pictured as several less attractive individuals) because of their facial characteristics. Finally, participants were instructed to think about this story while completing the JBT and try to overcome these

automatic processes. This intervention did not increase overall sensitivity, $t(871) = -1.06$, $p = .290$, $d = .07$, and did not reduce criterion biases, $t(871) = -1.33$, $p = .184$, $d = -.09$.

Blind Evaluation Instructions (Maximilian Primbs, Gijsbert Bijlstra, Ruddy Faure, and Johan C. Karremans). People have an automatic tendency to direct their attention to human faces (Langton et al., 2008), and to automatically infer traits and form evaluations based on faces' features, such as attractiveness (Olson & Marshuetz, 2005). These automatic tendencies affect responses, such as biases in hiring decisions in favor of physically attractive candidates (Axt et al., 2018). The present intervention aimed to disrupt such automatic tendencies by (a) making people aware of their existence, (b) highlighting their irrelevance, and (c) indicating what relevant information participants should focus on instead (the qualifications). This intervention did not increase overall sensitivity, $t(837) = -1.24$, $p = .214$, $d = .09$, and did not reduce criterion biases, $t(837) = -1.56$, $p = .120$, $d = -.11$.

Moral First Impressions (Neil Hester). Research shows that people care deeply about being morally good (Prentice et al., 2019) and form strong global impressions about others based on their moral goodness (Goodwin et al., 2014). In this intervention, participants read an evidence-based paragraph describing that relying on faces to form first impressions can result in inaccurate judgments of targets' traits and mental states. Then, participants were told that inaccurately relying on faces to form first impressions results in unfair, harmful outcomes (e.g., unfair criminal and death sentences). To the extent that participants formed moral convictions about not unfairly judging others based on their faces, they should be more motivated to self-monitor their preferences for attractive faces (Skitka et al., 2021). This intervention did not increase overall sensitivity, $t(811) = -0.21$, $p = .833$, $d = .01$, and did not reduce criterion biases, $t(811) = -1.55$, $p = .122$, $d = -.11$.

Counter-Stereotypical Exemplar Interventions

Vivid Narrative Exercise (Amy Hackney and Rachel Cook). Previous work shows that a vivid second-person narrative depicting a racially counter-stereotypical villain and hero can reduce implicit racial preferences (Lai et al., 2014; Marini et al., 2012). This intervention presented participants with a counter-stereotypic story that associated a physically attractive character with negative emotions (e.g., rejection) and encouraged emotional closeness with a physically unattractive character. The task instructions were modelled from typical imagined intergroup contact tasks (Crisp et al., 2009; Vezzali et al., 2012). This intervention did increase overall sensitivity, $t(819) = -2.00$, $p = .046$, $d = .14$ and did not reduce criterion biases, $t(819) = 0.39$, $p = .695$, $d = -.03$. Since the intervention increased sensitivity in Study 1, it was included in Study 2. In the second round, this intervention did not increase overall sensitivity, $t(827) = 0.71$, $p = .476$, $d = -.05$, and did not reduce criterion biases, $t(827) = -0.57$, $p = .568$, $d = -.04$.

Associative Learning Paradigm (Kao-Wei Chua and Jonathan B. Freeman). Past studies demonstrated that a brief associative learning paradigm was effective in mitigating biases regarding facial trustworthiness across several direct and indirect measures (Chua & Freeman, 2021). This intervention involved a training paradigm where academic competence co-occurred with facial attractiveness. In the learning phase, participants viewed 20 people paired with a short behavioral sentence. Some examples of behavioral sentences included: “Mentors struggling students in their spare time” (competence) and “Does everything at the last minute” (incompetence). Faces consisted of an equal mix of male and female White faces with a neutral expression. Half of the faces were rated as high in facial attractiveness and the other half were rated as low in physical attractiveness. Eight of ten individuals high in attractiveness were associated with behaviors that were low in academic competence (e.g., “scored low on the

SAT") and eight of ten of the individuals low in attractiveness were associated with behaviors high in academic competence (e.g., "made the honor roll every semester"). Participants viewed the 20 individuals four times each, resulting in 80 total learning trials.

This intervention did not increase overall sensitivity, $t(559) = 1.11$, $p = .267$, $d = -.10$, and did not reduce criterion biases, $t(559) = 1.31$, $p = .191$, $d = -.12$. Since this intervention had a much lower sample in Study 1 ($N = 161$, average N in other intervention conditions = 350), we included it in Study 2, where it *decreased* overall sensitivity, $t(820) = 3.14$, $p = .002$, $d = -.22$ and did not reduce criterion biases, $t(820) = -0.07$, $p = .941$, $d = .01$.

Morality-Competence Exercise (Xi Shen, Ming Ma, Danila Medvedev, and Emily G. Ritchie). Diagnosticity plays a critical role in person perception (Ferguson et al., 2019; Skowronski & Carlston, 1989). Recent work has found that only extreme and diagnostic information can reverse both explicit and implicit impressions formed from facial appearances (Shen et al., 2020; Shen & Ferguson, 2021), suggesting that diagnostic propositional information that counters facial information can override people's reliance on faces. This intervention asked participants to learn extreme counter-facial information that varied on morality or competence (Fiske et al., 2007), such that less attractive faces were described as displaying moral and competent behaviors while more attractive faces were described as displaying immoral and incompetent behaviors. All faces presented were novel and not used in the JBT. Examples of moral/competent behaviors included stories about helping one's neighbor by taking care of their house while they are away, or being a highly successful student and earning scholarships. Examples of immoral/incompetent behaviors included driving while under the influence, or being the lowest performing employee for several years in a row. Participants were then challenged to consider this information as more informative than facial information.

This intervention did not increase overall sensitivity, $t(807) = -0.6$, $p = .551$, $d = .04$, and did not reduce criterion biases, $t(807) = -0.76$, $p = .450$, $d = -.05$.

Like-Dislike Writing (Samantha M. Stevens, Stephen Anderson, and Roger Beaty). Exposure to counter-stereotypic exemplars can reduce implicit bias (Dasgupta & Greenwald, 2001), and counter-stereotypic mental imagery can reduce implicit stereotypes (Blair et al., 2001). This intervention combines creative mindset ideas (Sassenberg & Moskowitz, 2005) with counter-stereotypic thinking to potential judgment biases concerning attractiveness. Participants completed four writing exercises (two with men targets, two with women targets; all White with neutral expressions), with each gender having one more attractive face and one less attractive face. For less attractive targets, participants listed reasons that their friends would like the target. For more attractive targets, participants listed reasons that their friends would dislike the target. We asked that participants to use their imagination and be as creative as possible in their responses. All four writing exercises were completed on the same page, and participants were required to spend between four and six minutes on the task. This intervention did not increase overall sensitivity, $t(718) = 1.18$, $p = .238$, $d = -.09$, and did not reduce criterion biases, $t(618.92) = -0.51$, $p = .608$, $d = -.04$.

Imagined Contact (Sandro Rubichi, Loris Vezzali, and Veronica Margherita Cocco). According to Crisp and Turner (2012), direct intergroup contact is not necessary to improve intergroup relations. Rather, the simulation of an intergroup interaction is sufficient to reduce ingroup bias. This intervention applied imagined contact to attractiveness biases, by focusing on the role played by competence stereotypes (Fiske et al., 2002). Participants were asked to imagine contact with a competent colleague to promote the relevance of competence (rather than attractiveness) when evaluating others, and to imagine that this interpersonal contact was

successful (Vezzali et al., 2015). This intervention did not increase overall sensitivity, $t(645) = 0.97$, $p = .330$, $d = -.08$, and did not reduce criterion biases, $t(645) = 1.13$, $p = .258$, $d = .09$.

Norm Information (Stephanie Johnson and Elsa Chan). Although attractiveness is generally valued, individuals also hold negative stereotypes about attractive people and negative views can become more pronounced when attractiveness does not seem relevant to the context (Johnson & Chan, 2019). When such ambiguity is present, decisions can be influenced by norms that are not only descriptive of the qualities of the current options but also prescriptive of what the options should be (Samuelson & Zeckhauser, 1988). In this intervention, participants were presented with photos of the current members of the academic honor society, and these members had faces that were rated as relatively unattractive. This intervention did not increase overall sensitivity, $t(872) = -0.14$, $p = .892$, $d = .01$, and did not reduce criterion biases, $t(872) = 0.61$, $p = .545$, $d = .04$.

Attractive = Harmful Movie (Yalda Daryani). One of the prevailing viewpoints in moral psychology is that attractiveness bias exists because beautiful or eye-catching individuals trigger the moral foundation of care (Haidt, 2012). Given that "harm" is the opposite of "care" (Graham et al., 2013), this intervention proposed that when people are exposed to scenes in which beautiful people are engaged in harmful activities, the care foundation would not be triggered and participants would learn that being attractive does not equal being harmless. In this intervention, participants watched two short film clips (total time = 4.3 minutes) in which the main character is an attractive person and engages in harmful acts (e.g., killing civilians). This intervention did not increase overall sensitivity, $t(763) = 0.87$, $p = .385$, $d = -.06$, and did not reduce criterion biases, $t(763) = -0.53$, $p = .600$, $d = -.04$.

General Self-Reflection Interventions

Cultural Self-Awareness (Chieh Lu and Yen-Ping Chang). Cultural self-awareness refers to individuals' awareness of culture's influence on their self (Lu & Wan, 2018). Individuals with high cultural self-awareness apply culturally-constituted evaluations in a more conscious manner and are more reserved in evaluating others. Previous research has shown that in certain situations, cultural self-awareness can reduce prejudicial attitudes (Lu, Lee, & Wan, 2023). This intervention asked participants to write about how culture(s) they were immersed in early in life have shaped who they are today. This intervention did not increase overall sensitivity, $t(670) = -0.01$, $p = .993$, $d < .01$, and did not reduce criterion biases, $t(670) = -1.03$, $p = .305$, $d = -.08$.

Mindfulness Exercise (Denise Zheng and Kirk Warren Brown). Previous research showed that a 10-minute mindfulness meditation decreased implicit biases toward racial out-groups (Lueke & Gibson, 2015) and decreased discrimination in an economic trust game (Lueke & Gibson, 2016). Based on these findings, this intervention had participants listen to a 7-minute mindfulness audio recording. The intervention instructed participants to use their breath as the anchor to foster greater receptive attention to and awareness of the psychological and somatic experiences that arose during practice. Using a type of focused attention (FA) mindfulness practice (Lutz et al., 2015), participants were instructed to acknowledge when their mind has wandered away from the anchor during the task and then bring their attention back to their breath. This intervention did increase overall sensitivity, $t(635) = -2.18$, $p = .030$, $d = .18$, and did not reduce criterion biases, $t(635) = -1.01$, $p = .314$, $d = -.08$.

While this intervention increased sensitivity in Study 1, it was not initially included in Study 2 due to an undetected error in Study 1 analyses. As a result, we conducted a separate replication on Prolific, randomly assigning participants to the intervention or a control condition

(total eligible $N = 811$). Again, the control condition showed a robust criterion bias ($d = .30$) and above-chance levels of JBT accuracy (67.2%). The intervention did not increase overall sensitivity, $t(809) = 0.44$, $p = .658$, $d = -.05$ and did not reduce criterion biases, $t(809) = -0.72$, $p = .475$, $d = -.03$.

Similarity-Attraction Paradigm (Lusine Grigoryan and M. Isabelle Weiβflog). The similarity-attraction paradigm (Byrne, 1971) postulates that people tend to like others who are similar to them. In a related line of research, studies on crossed categorization show that people have more positive attitudes towards targets that share at least one form of group membership with them (Crisp & Hewstone, 2007). This intervention was designed to increase perceived similarity between participants and less physically attractive targets. Participants were given information about targets' group memberships along the dimensions of religion, political affiliation, nationality, and occupation, which have a relatively strong effect on attitudes (Grigoryan et al., 2022). Participants first reported their own group membership on each dimension. This information was then used to manipulate the number of shared group memberships between participants and targets: 3-4 shared group memberships were presented for less attractive targets, and 0-1 shared group memberships were presented for more attractive targets. Participants viewed 12 applicants and selected all characteristics they shared with each target. All faces presented were novel and not used in the JBT. This intervention did not increase overall sensitivity, $t(809) = 0.03$, $p = .978$, $d < .01$, and did not reduce criterion biases, $t(809) = 0.02$, $p = .985$, $d < .01$.

Evaluation Criteria (Provided) Interventions

Two-Out-of-Three Rule (Bastian Jaeger and Anthony M. Evans). Decades of research have shown that people often use mental short cuts, such as relying on easily accessible cues,

when making decisions (Gigerenzer & Goldstein, 1996; Tversky & Kahneman, 1974). Faces attract attention (Theeuwes & Van der Stigchel, 2006), attractiveness judgments can be formed rapidly (Ritchie et al., 2017), and reliance on face judgments is relatively effortless (Jaeger et al., 2019). This may explain the widespread effects of face judgments on decision-making. In this intervention, participants were informed that decisions are often biased by irrelevant characteristics that are visible from a person's photo. Participants were instructed to apply a simple decision rule to avoid this bias. Specifically, participants were told that applicants should be admitted if they fulfilled at least two of three criteria: a science GPA above 3.3, excellent recommendation letters, and an interview score above 80. Participants completed three practice rounds where they received feedback on whether they applied the rule successfully. All faces and qualification values presented were novel and not used in the JBT.

In Study 1, this intervention did increase overall sensitivity, $t(809.12) = -18.93$, $p < .001$, $d = 1.29$, and also reduced criterion biases, $t(683.33) = -2.40$, $p = .011$, $d = -.17$. Since this intervention both reduced criterion bias and increased sensitivity in Study 1, it was included in Study 2. In Study 2, the intervention again increased overall sensitivity, $t(800.98) = -19.07$, $p < .001$, $d = 1.30$, and reduced criterion biases, $t(706.81) = -4.13$, $p < .001$, $d = -.29$.

Criteria Reinforcement Exercise (Kate M. Turetsky). One mechanism for discrimination is that people apply different standards when evaluating members of favored and disfavored social groups (e.g., Hodson et al., 2002; Uhlmann & Cohen, 2005). Setting clear, objective, and universal criteria for evaluation before considering candidates can reduce discrimination by reducing the flexibility evaluators have to adjust their standards (Quinn, 2020). This intervention provided participants with the averages of the candidate pool's science and humanity GPAs, recommendation letter ratings, and interview scores. Participants were told to accept the

candidates who were above average overall when using these four components. The intervention had an eight-trial practice round with novel applications that provided feedback on each trial. This intervention did increase overall sensitivity, $t(641.62) = 14.89$, $p < .001$, $d = 1.13$ and reduced criterion biases, $t(703.22) = -2.59$, $p = .010$, $d = -.19$. Since this intervention both reduced criterion bias and increased sensitivity in Study 1, it was included in Study 2. In Study 2, this intervention again increased overall sensitivity, $t(747.84) = -13.26$, $p < .001$, $d = .90$, and reduced criterion biases, $t(763.34) = -4.41$, $p < .001$, $d = -.31$.

Propositional-Statistical Learning (Xin Yang and Arin Korkmaz). Propositions, or statements about relations between concepts, are effective in shaping social learning (De Houwer, 2014) and especially effective in updating seemingly robust face-based impressions (Shen et al., 2020). Attractiveness bias could be mitigated by a propositional statement about qualifications (e.g., "Qualified candidates have X…"), followed by statistical learning to establish that proposition's reliability and decrease the reliance on physical attractiveness. Participants read a description of qualifications from previous, excellent applicants and were told to use these same qualification benchmarks to create their own excellent team. Specifically, participants were given a rule consisting of pre-determined cut-offs based on some of the JBT's criteria, asking participants to accept applicants if they had Science GPA > 3.5 and Interview Scores > 75. In addition, participants were instructed to conditionally accept those who were reasonably close to those cut-offs and had a Humanities GPA > 3.5. Finally, participants reviewed 16 novel candidates in a JBT-like paradigm and predicted whether they were selected based on the proposition provided, receiving feedback after each decision. In Study 1, this intervention actually decreased overall sensitivity, $t(684.21) = 3.23$, $p = .001$, $d = -.23$, but did reduce criterion biases, $t(685.62) = -2.44$, $p = .015$, $d = -.17$. Since this intervention reduced

biases in response criterion in Study 1, it was included in Study 2. In Study 2, the intervention again decreased overall sensitivity, $t(746.28) = 6.82$, $p < .001$, $d = -.47$ and reduced criterion biases, $t(715.03) = -4.84$, $p < .001$, $d = -.34$.

Evaluation Criteria (Self-Determined) Interventions

Minimal Threshold Creation (Ariella Kristal). Precommitment has been shown to be effective at helping individuals make decisions that better align with their values (Milkman et al., 2008; Thaler & Benartzi, 2004). While previous research has shown that precommitting to decision criteria in advance can prevent people from changing selection criteria to justify selecting a candidate they wanted to hire (Uhlmann & Cohen, 2005), the current intervention sought to test whether precommitment can combat other selection biases. Participants completed a practice round of the JBT using novel faces and applications, then read about the benefits of precommitment before being asked to generate minimum thresholds for each of the four qualification indicators. Participants were then told they should apply this rule in the JBT and accept the candidate only if they exceeded each self-created threshold. This intervention did not increase overall sensitivity, $t(723) = 0.91$, $p = .365$, $d = -.07$, and did not reduce criterion biases, $t(723) = -1.17$, $p = .242$, $d = -.09$.

Single-Criterion Exercise (Balbir Singh and Joshua Correll). The aversive racism framework suggests that bias occurs when qualifications are ambiguous (Dovidio & Gaertner, 2000; Hodson et al., 2002). Ambiguous qualifications allow perceivers to flexibly weigh particular dimensions in order to justify biased decisions (Norton et al., 2004, 2006). This intervention sought to discourage differential weighting by encouraging participants to focus on a single qualification and to use that qualification consistently. Despite the contest study rewarding interventions that impacted both criterion bias and sensitivity, this intervention was

primarily directed at reducing criterion biases. Participants began with a training block, responding to 16 novel applicants presented with qualifications but not faces. After each of the first eight training trials, participants were asked to write a few sentences justifying their decisions. Participants were then asked to write down the qualification they thought was most relevant to differentiating between more and less qualified candidates and the one they thought was the least relevant. Next, they had to generate a rule based on the single qualification deemed most important, thus minimizing ambiguity. They then practiced applying their rule with an additional eight practice trials. Finally, prior to starting the JBT, participants were reminded that they should consistently apply their rule. This intervention did not increase overall sensitivity, $t(698) = -0.28$, $p = .783$, $d = .02$, but did reduce criterion biases, $t(698) = -2.46$, $p = .014$, $d = -.19$. Since this intervention reduced biases in response criterion in Study 1, it was included in Study 2. In Study 2, this intervention now reduced overall sensitivity, $t(790) = 4.41$, $p < .001$, $d = -.31$, and again reduced criterion biases, $t(790) = -2.24$, $p = .026$, $d = -.16$.

Criteria from Mean Values (Erika L. Kirgios, Linda W. Chang, and Edward H. Chang). Discrimination is more common under conditions of ambiguity (Hodson et al., 2002; Uhlmann & Cohen, 2005; Dovidio & Gaertner, 2000). One potential strategy to reduce ambiguity in hiring decisions—and mitigate discriminatory decision-making—is to ask evaluators to pre-commit to decision rules (Uhlmann & Cohen, 2005). This intervention asked participants to self-generate a decision rule that specified the conditions under which they would "Accept" or "Reject" candidates. Specifically, participants were instructed to pay attention to the range of values present in the JBT's encoding phase. After the encoding phase, participants were first asked to estimate the overall average value of each of the four qualifications. Then, they were also told to write out a rule to guide their decision-making; participants were given an example

rule of only accepting candidates who were above the perceived average on at least three dimensions. Participants then filled in their own decision rule by completing the phrase "I will only accept candidates if…". Before starting the JBT, participants were shown their decision rule and asked to apply this rule to all candidates. This intervention did not increase overall sensitivity, $t(693) = 0.38$, $p = .705$, $d = -.03$, but did reduce criterion biases, $t(693) = -2.06$, $p = .039$, $d = -.16$. Since this intervention reduced criterion bias in Study 1, it was included in Study 2. In Study 2, this intervention again did not increase overall sensitivity, $t(742.52) = -0.64$, $p = .525$, $d = -.04$, but did reduce criterion biases, $t(785.37) = -2.86$, $p = .004$, $d = -.20$.

Criteria from Min-Max Values (Jennifer Steele, Julia Sebastien, and Jennifer Sedgwick). Biases based on group membership have been found in a number of hiring contexts (Biernat & Kobrynowicz, 1997; Goldin & Rouse, 2000), and initial research suggests that creating evaluation criteria prior to selecting candidates has the potential to decrease bias (Uhlmann & Cohen, 2005). In this intervention, participants were introduced to the JBT and were then told that people make less biased decisions when they outline criteria before selecting applicants. Participants were reminded of the four types of information that they would receive about each applicant (e.g., Science GPA) and were provided with the range of scores that these applicants might have (e.g., 2.9-3.9). They were then encouraged to decide on the ideal threshold that applicants would need to meet in order to be selected. To increase accountability, participants were asked to outline their criteria, both online and on a sheet of paper prior to starting the task (the paper would remain visible throughout the task).. This intervention did not increase overall sensitivity, $t(699) = -1.40$, $p = .161$, $d = .11$, but did reduce criterion biases, $t(699) = -2.04$, $p = .042$, $d = -.16$. Since this intervention reduced criterion bias in Study 1, it was

included in Study 2. In Study 2, the intervention did not increase overall sensitivity, $t(804) = 1.55$, $p = .122$, $d = -.11$ and did not reduce criterion biases, $t(804) = --0.57$, $p = .567$, $d = -.04$.

Personal Information Avoidance Rule (Sean Fath). Cues to engage in reflective thinking (Kahneman, 2011) may encourage selective attention to useful information. For instance, evaluators are less likely to choose to view potentially biasing information about a target after an intervention prompting reflection on the potential for bias associated with receipt of such information (Fath et al., 2022). This intervention prompted participants to approach the evaluation task in a reflective mindset. Participants were presented with a summary of the information they would receive about the applicants, organized into "personal information" (i.e., photo) and "qualification information" (i.e., Science GPA, Humanities GPA, Letters of recommendation, and Interview score). Next, they were asked to indicate the information they thought they should focus on in order to provide an unbiased judgment. This intervention did not increase overall sensitivity, $t(792.76) = 0.41$, $p = .686$, $d = -.03$, and did not reduce criterion biases, $t(815) = -0.41$, $p = .684$, $d = -.03$.

Discussion

Study 1 used a contest design to test 30 interventions submitted by social scientists to reduce discrimination in a task known to produce favoritism in judgment based on physical attractiveness. In a series of well-powered tests, the first round of the contest found that two interventions reduced relative biases in response criterion *and* increased overall sensitivity, whereas four interventions only reduced criterion biases and three interventions only increased overall sensitivity (one additional intervention produced marginally significant results for increasing sensitivity). Study 2 then tested these nine interventions – as well as the intervention

that produced the marginally significant results and one final intervention with a notable small Study 1 sample size – on a new sample source.

In this follow-up study, one intervention now reduced sensitivity and four interventions failed to reliably impact either sensitivity or criterion bias. Two interventions (Two Out of Three Rule, Criteria Reinforcement Exercise) showed greater levels of effectiveness by both reliably increasing sensitivity and reducing criterion biases, while two interventions showed more moderate effects by either only increasing sensitivity (Trial-and-Error Feedback) or only reducing criterion biases (Criteria from Mean Values). Finally, two interventions reduced biases in response criterion for more versus less physically attractive applicants but did so while also lowering sensitivity and thereby increasing the total amount of unfair treatment across all applicants (i.e., overall instances of accepting less qualified applicants and rejecting more qualified applicants). As a result, it is not clear that these interventions led to a more desirable behavior, as increases in judgment errors led to a greater number of applicants being the recipient of unfair treatment. For this reason, these two interventions (Propositional-Statistical Learning and Single-Criterion Exercise) were not retained for Studies 3-4.

Study 3

In Study 3, we explored the generalizability of the four interventions that reduced bias, increased sensitivity, or impacted both outcomes in Studies 1-2. One limitation of Studies 1-2 is that the JBT only focused on a single form of social bias (physical attractiveness), which overlooks the fact that people possess multiple social identities and as a result may be susceptible to several forms of discrimination that operate simultaneously. To address this issue, Study 3 used a version of the JBT that has been shown to produce two simultaneous (and independent) forms of judgment bias (Axt et al., 2019). Specifically, participants viewed

applicants that were either more or less physically attractive and were members of a political ingroup or outgroup (i.e., Democrats or Republicans). Through using these same interventions in a new context, Study 3 sought to identify interventions that address multiple forms of discrimination and highlight strategies that can instil a more domain-general ability to ignore potentially biasing social information.

Methods

Participants were recruited from Prolific. We only included the four interventions that either reliably reduced criterion biases in Studies 1-2 (Criteria from Mean Values), increased sensitivity (Trial-and-Error Feedback) or both reduced biases and increased sensitivity (Two Out of Three Rule, Criteria Reinforcement exercise).[5] We also excluded the two interventions (Propositional-Statistical Learning and Single-Criterion Exercise) that reduced discrimination by lowering criterion bias but also reliably decreased sensitivity (i.e., increased judgment errors) as they ultimately led to more unfair treatment.

We restricted the study to Prolific participants who were residents of the United States and who reported being either Republican or Democrat when first registering for Prolific. A total of 2425 participants completed the study. We excluded participants based on the same JBT criteria used for the second study (n = 121) and also screened out participants who did not explicitly report identifying as Republican or Democrat when asked during the study session (n = 298)[6]. The resulting sample was N = 2006 (M_{age} = 35.3, SD_{age} = 12.7, 78.1% White, 39.1% female). We collected an average of 401 participants per condition, which provided more than 80%

[5] The Trial-and-Error Feedback intervention was only marginally significant in Study 1, but significantly increased sensitivity in Study 2.
[6] When including all participants with full JBT data in Study 3, no conclusions changed.

power to detect an effect as small as d = .20 between each intervention and the control condition.

Participants followed the same procedure outlined in Study 1, with the exception that they completed a modified version of the JBT (Axt et al., 2019). Participants were randomly assigned to complete one of the four interventions – Criteria from Mean Values, Criteria Reinforcement Exercise, Trial-and-Error Feedback, and Two-Out-of-Three Rule – or to a control condition, and then completed the JBT. In this dual-bias version of the JBT, the 64 applicants were presented with information related to physical attractiveness (less vs more; communicated by a face) and political party affiliation (Democrat vs. Republican; communicated by a party logo), as well as qualification information that made the applicant more or less qualified using the same procedure as Studies 1-2.

After completing the JBT, participants answered self-report items about their desired and perceived performance on the task. Lastly, participants completed a 7-item demographic questionnaire and an attention check item.

Results

For analyses, judgments were recoded to focus on whether the applicant was a political ingroup or outgroup member. In Study 3, the control condition showed above-chance, moderate levels of overall accuracy (66.1%), leaving significant room for interventions to improve accuracy/sensitivity. A 2 (Physical attractiveness: more vs. less) by 2 (Political group status: ingroup vs. outgroup) ANOVA on response criterion values revealed two simultaneous biases in criterion in the control condition; a main effect of physical attractiveness (η_p^2 = .048, p < .001) showing that on average, more physically attractive applicants had a lower response criterion (M = -.04, SD = .59) than less physically attractive applicants (M = .04, SD = .58), as well as a main

effect of political ingroup status (η_p^2 = .172, p < .001) showing that on average political ingroup members received a lower response criterion (*M* = -.18, *SD* = .58) than political outgroup members (*M* = .18, *SD* = .58). There was no interaction between attractiveness and political ingroup status in response criterion (η_p^2 = .004, p = .168; see online supplement for full reporting).

Given that the control condition successfully produced two separate criterion biases, we then ran a series of 2 (Applicant attractiveness) by 2 (Applicant political ingroup status) by 2 (Condition: Control versus Intervention) ANOVAs on criterion values. Here, an interaction between applicant attractiveness and condition would indicate that the intervention impacted criterion biases regarding physical attractiveness, and an interaction between applicant political ingroup status and condition would indicate that the intervention impacted criterion biases regarding political ingroup status. Finally, to examine the effects of each intervention on sensitivity, we performed a *t*-test on the difference in overall JBT sensitivity for each intervention relative to the control condition. See Table 4 for sample size as well as descriptive statistics for overall sensitivity and response criterion for more attractive and less attractive ingroup and outgroup members. Figure 3 displays results comparing each intervention relative to control on overall JBT sensitivity, and Figure 4 displays results (converted to a Cohen's *d* effect size) for each intervention relative to control on criterion biases based on applicants' physical attractiveness or political ingroup status.

Trial-and-Error Feedback. This intervention increased overall sensitivity, t(816.87) = 2.49, p = .013, d = .17. For response criterion, the interaction between political ingroup and condition was not significant (F(1, 858) = .883, p = .348, η_p^2 = .001), such that participants in the intervention still showed an effect of political ingroup status on criterion (η_p^2 = .142) that did

not reliably differ from the control condition. The interaction between attractiveness and condition was also not significant ($F(1, 858) = .028$, $p = .867$, $\eta_p^2 = .000$), such that participants in the intervention still showed an effect of attractiveness on criterion ($\eta_p^2 = .059$) that did not differ from the control condition (see online supplement for full reporting).

Two-Out-of-Three Rule. This intervention did increase overall sensitivity, $t(745.75) = 17.94$, $p < .001$, $d = 1.21$. For response criterion, the interaction between political ingroup and condition was significant ($F(1, 893) = 45.5$, $p < .001$, $\eta_p^2 = .048$), such that among participants in the intervention, there was a present but reliably smaller effect of political ingroup status on criterion ($\eta_p^2 = .028$). The interaction between attractiveness and condition was also significant ($F(1, 893) = 12.46$, $p < .001$, $\eta_p^2 = .014$), such that among participants in the intervention, there was no main effect of attractiveness on criterion ($\eta_p^2 = .001$; see online supplement for full reporting).

Table 4.
Means and Standard Deviations of JBT Outcomes for Each Condition in Study 3

		More Attractive Criterion		Less Attractive Criterion	
Condition	Sensitivity	Ingroup Members	Outgroup Members	Ingroup Members	Outgroup Members
Control (*N* = 464)	0.94 (.59)	-0.23 (.59)	0.15 (.58)	-0.13 (.57)	0.21 (.59)
Trial-and-Error Feedback (*N* = 388)	1.05 (.60)	-0.24 (.54)	0.08 (.56)	-0.15 (.57)	0.16 (.56)
Two-Out-of-Three Rule (*N* = 428)	1.84 (.85)	.06 (.36)	0.13 (.37)	0.07 (.40)	0.14 (.38)
Criteria Reinforcement Exercise (*N* = 380)	1.63 (.68)	0.25 (.49)	0.31 (.45)	0.27 (.47)	0.31 (.45)
Criteria from Mean Values (*N* = 351)	0.94 (.58)	-0.01 (.54)	0.12 (.52)	0.02 (.53)	0.23 (.52)

Figure 3. Difference in Sensitivity with the Control Condition for each Intervention in Study 3.

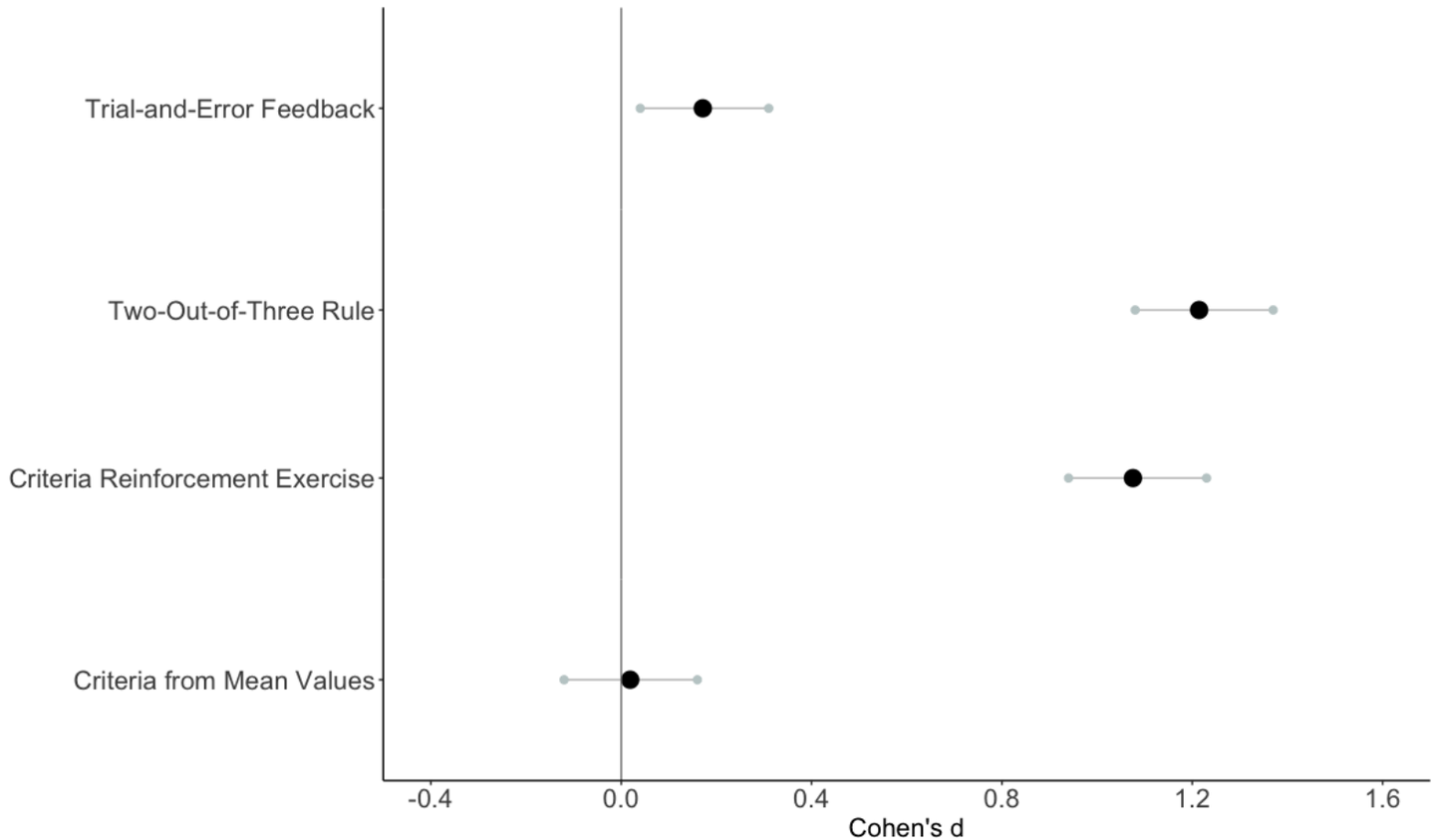
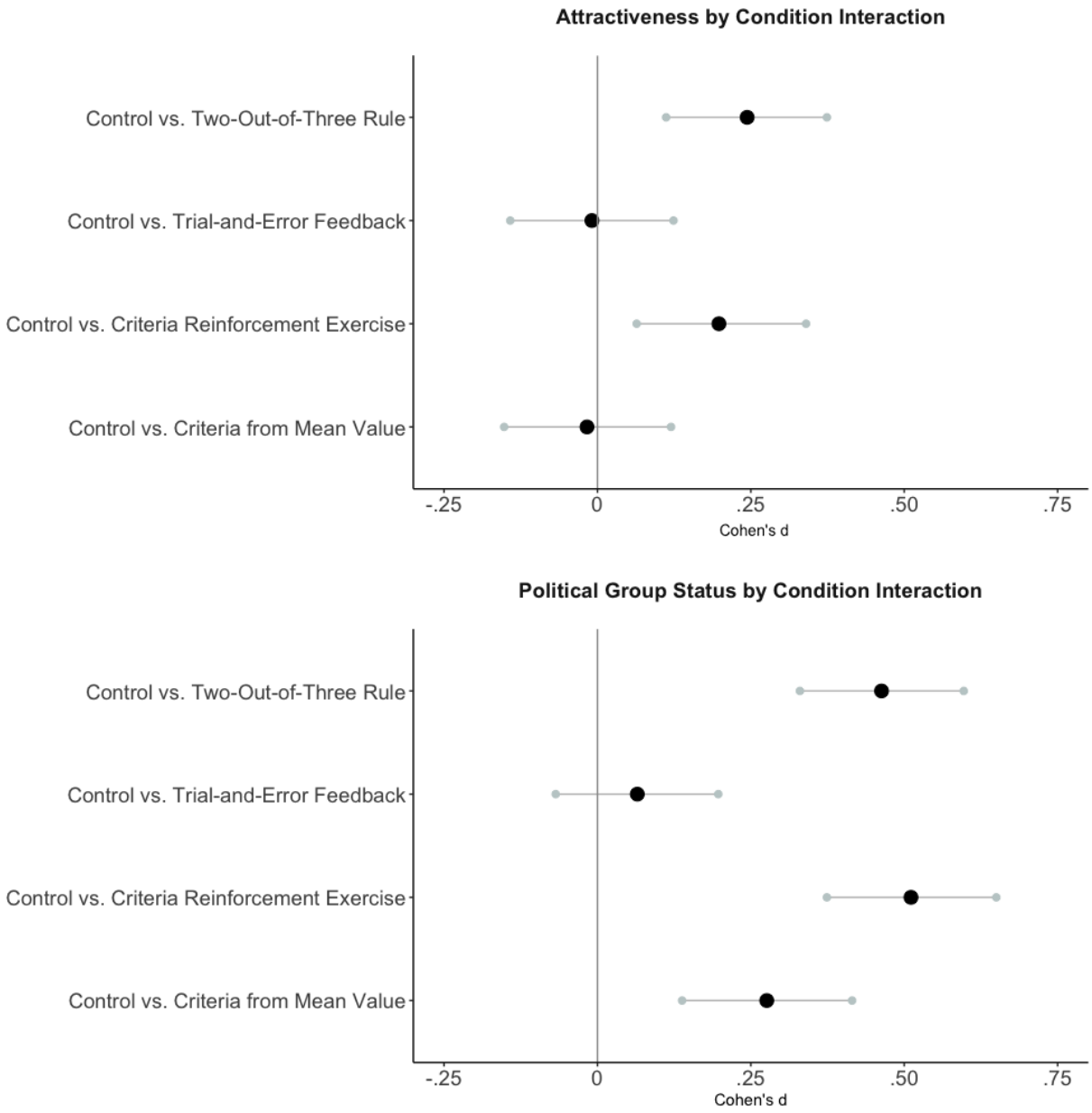


Figure 4. Effect Size of the Interactions between Criterion biases and Condition for Study 3.



Criteria Reinforcement Exercise. This intervention did increase overall sensitivity, $t(755.26) = 15.42$, $p < .001$, $d = 1.08$. For response criterion, the interaction between political ingroup and condition was significant ($F(1, 846) = 51.734$, $p < .001$, $\eta_p^2 = .058$), such that among participants in the intervention, there was a present but smaller effect of political ingroup status on criterion ($\eta_p^2 = .026$). The interaction between attractiveness and condition was also significant ($F(1, 846) = 8.276$, $p = .004$, $\eta_p^2 = .010$), such that among participants in the intervention, there was no main effect of attractiveness on criterion ($\eta_p^2 = .003$; see online supplement for full reporting).

Criteria from Mean Values. As in Studies 1-2, this intervention did not increase overall sensitivity, $t(762.63) = .27$, $p = .790$, $d = .02$. For response criterion, the interaction between political ingroup and condition was significant ($F(1, 816) = 15.190$, $p < .001$, $\eta_p^2 = .018$), such that among participants in the intervention, there was a present but smaller effect of political ingroup status on criterion ($\eta_p^2 = .085$). However, the interaction between attractiveness and condition was not significant ($F(1, 816) = .44$, $p = .834$, $\eta_p^2 = .000$), such that participants in the intervention still showed an effect of attractiveness on criterion ($\eta_p^2 = .056$) that did not reliably differ from the control condition (see online supplement for full reporting).

Discussion

Study 3 used a JBT that, in control conditions, produced two simultaneous and independent biases based on physical attractiveness and political ingroup status. Two interventions (Two-Out-of-Three Rule and Criteria Reinforcement Exercise) produced similar effects as Studies 1-2 in that each strategy both improved sensitivity and now reduced two forms of criterion bias. Another intervention (Criteria from Mean Values), was only effective at

reducing criterion biases based on political ingroup affiliation, leaving criterion biases based on physical attractiveness unchanged. This result is a contrast with Studies 1-2, where the intervention reduced criterion biases based on physical attractiveness. As in Studies 1-2, the Criteria from Mean Values intervention did not impact overall sensitivity. Finally, another intervention (Trial-and-Error feedback) showed results consistent with Studies 1-2 in that the intervention only increased sensitivity. One potential limitation of Studies 1-3 is the reliance on convenience samples. It is possible that participants with hiring experience may be less susceptible to these judgment biases (e.g., Lehman & Nisbett, 1990) or less responsive to any of the proposed interventions due to greater confidence in their ability to evaluate others (e.g., Fisher et al., 2020). To explore this possibility, Study 4 tested the effectiveness of the four Study 3 interventions on participants who reported having experience with hiring employees.

Study 4

Methods

Participants were recruited from Prolific, and only participants who responded "yes" to the pre-screen item "Do you have any experience in making hiring decisions (i.e. have you been responsible for hiring job candidates)?" were eligible to view the study. We added an additional screener item within the study session where participants had to report that they "had experience in hiring employees." Despite 2821 participants completing the study, only 2318 passed the pre-screen item about hiring experience included in the actual study, leaving a final sample of 2162 (M_{age} = 40.4, SD_{age} = 12.3, 71.3% White, 45.3% female) who in addition passed an attention check item and had the same JBT performance criteria as used in Studies 1-2. Eligibility criteria, sample sizes, methods and analyses were pre-registered at https://osf.io/gbpqa/. Conclusions do

not change when including all participants with self-reported hiring experience (i.e., removing exclusions based on JBT criteria).

In our pre-registered analyses concerning the effectiveness of each intervention, we used one-tailed analyses.[7] Past work highlights how, when pre-registered, one-tailed analyses can be an effective means of increasing statistical power and controlling error rates (Hales, 2023). Specifically, we tested whether, relative to the control condition, each intervention 1) reduced criterion biases and 2) increased overall sensitivity. Though the four interventions used in Study 4 did not each consistently produce these two effects in Studies 1-3, we still pre-registered these comparisons to maximize similarity across analyses.

Participants followed the same procedure outlined in Study 2 but were assigned to one of the four interventions used in Study 3. As in Studies 1-2, the JBT only included a manipulation of target physical attractiveness. Aside from including only four interventions, there were two other changes from Study 2. The first was the addition of the two items about whether the participant had hiring experience and how many years of experience they had hiring employees (this item was not included in primary pre-registered analyses but is available in the online dataset). Overall, 83.3% of participants reported having more than one year of hiring experience.

The second change was the use of a weighted random assignment to conditions based off meta-analytic effect sizes from Studies 1-2, such that the Two-out-of-Three condition and Criteria Reinforcement Exercise condition had approximately half as many participants as the Control condition or two other intervention conditions (see online supplement for more detail

[7] All the conclusions hold when using two-tailed analyses, with one exception: for the Trial-and-Error feedback intervention, the two-tailed test found only a marginally significant reduction in criterion bias, $t(1145.12) = 1.90$, $p = .058$, $d = -.11$.

about power analysis calculations). Using one-tailed analyses and effect sizes calculated from meta-analyzing the results of Studies 1-2, the sample sizes obtained in Study 4 allowed for a minimum statistical power of 84.4% and an average power of 92.4% for tests of each intervention's effectiveness relative to the control condition (see pre-registration file for more details about anticipated effect sizes and target sample sizes).

Results

Table 5 reports sample sizes for each condition as well as means, standard deviations, and (two-tailed) within-subjects *t*-tests comparing criterion for more versus less physically attractive applicants. There tests were two-tailed to align with our pre-registration.

Trial-and-Error Feedback. This intervention increased overall sensitivity, $t(1136.15) = 3.92$, $p < .001$, $d = .23$, and reduced criterion biases, $t(1145.12) = 1.90$, $p = .029$, $d = -.11$. This latter result – a significant reduction in criterion bias – was not in our pre-registered set of predictions, nor was it observed in Studies 1-3.

Two-Out-of-Three Rule. This intervention increased overall sensitivity, $t(432.02) = 19.17$, $p < .001$, $d = 1.46$, and reduced criterion biases, $t(738.43) = 3.23$, $p < .001$, $d = -.22$.

Criteria Reinforcement Exercise. This intervention increased overall sensitivity, $t(317.22) = -13.60$, $p < .001$, $d = 1.13$, and reduced criterion biases, $t(633.71) = 2.58$, $p = .005$, $d = -.18$.

Criteria from Mean Values. This intervention did not increase overall sensitivity, $t(1100) = 1.63$, $p = .052$, $d = .10$, though the effect was marginally significant. The intervention did reliably reduce criterion biases, $t(1098.02) = 2.17$, $p = .015$, $d = -.13$.

Table 5.

Sample sizes, Means, Standard Deviations, and Within-Subjects *t*-tests of JBT Outcomes for Each Condition in Study 4.

Condition	Sensitivity	More Attractive *c*	Less Attractive *c*	Criterion *t*-test
Control (*N* = 601)	.96 (.53)	-.09 (.43)	.01 (.42)	$t(600) = 5.70$, $p < .001$, $d = .23$
Trial-and-Error Feedback (*N* = 561)	1.09 (.57)	-.06 (.42)	-.003 (.41)	$t(560) = 3.71$, $p < .001$, $d = .16$
Two-Out-of-Three Rule (*N* = 273)	1.84 (.67)	.14 (.34)	.16 (.36)	$t(272) = 1.07$, $p = .287$, $d = .06$
Criteria Reinforcement Exercise (*N* = 226)	1.69 (.73)	.32 (.50)	.35 (.49)	$t(225) = 1.92$, $p = .057$, $d = .13$
Criteria from Mean Values (*N* = 501)	1.01 (.51)	.05 (.49)	.10 (.48)	$t(500) = 3.21$, $p = .001$, $d = .14$

Discussion

In a sample of participants with self-reported hiring experience, those completing the JBT with no intervention showed biases in criterion that favored more over less physically attractive applicants.[8] In addition, three of the interventions in Study 4 were able to change both criterion bias and sensitivity (Trial-and-Error Feedback, Two-out-of-Three Rule, Criteria Reinforcement Exercise). Another intervention only reduced criterion bias (Criteria from Mean Values), though this intervention also produced a marginally significant increase in sensitivity. The Study 4 results then suggest that the effects of these interventions are not limited to novice participants. Rather, results seem to extend to decision-makers with self-reported experience of evaluating others within a professional context. However, one weakness of Study 4 is its reliance on self-reported experience rather than directly sampling participants with more known experience (e.g., by recruiting at a conference for human resources professionals).

General Discussion

Thirty interventions aiming to reduce discrimination in social judgment were tested over four rounds of a research contest that involved over 20,000 participants. The interventions were submitted by teams of researchers, with 97% of teams having at least one member with a Ph.D. in psychology or a related field (e.g., organizational behavior), and 90% of teams having at least one member that previously published work on stereotypes, prejudice, or discrimination. In the first round of the contest, nine of the 30 interventions reliably reduced decision bias and/or decision noise in our chosen outcome – the JBT – relative to a control condition.

[8] The online supplement reports a pilot study using the same eligibility criteria as Study 4 (N = 266) that replicated the criterion bias on the JBT among control participants (d = .23, p < .001).

The second round of the contest included a direct replication of these nine interventions (plus two additional interventions) using a different sample source. Results revealed that two interventions were effective at reducing both bias and noise, while two interventions only reduced bias or noise. A third round found that these four interventions retained some level of effectiveness in terms of changing criterion bias and/or increasing overall sensitivity when using a JBT that produced two simultaneous and independent forms of discrimination (i.e., based on physical attractiveness and political ingroup identity), and a final round found that each of the interventions reduced either bias, noise, or both outcomes among a sample of participants with self-reported hiring experience.

However, while four interventions were consistently effective at reducing decision bias, decision noise, or both outcomes on the JBT, the interventions were far from equally effective. Meta-analyses of Studies 1, 2 and 4 (those using a JBT dealing solely with attractiveness-based discrimination) found that two interventions, Criteria Reinforcement Exercise and Two-Out-of-Three Rule, produced large effects on sensitivity (Criteria Reinforcement $d = 1.04$, $p < .001$; Two-Out-of-Three $d = 1.35$, $p < .001$) as well as moderate but robust effects on criterion bias (Criteria Reinforcement $d = -.23$, $p < .001$; Two-Out-of-Three $d = -.23$, $p < .001$). The other two interventions, Trial-and-Error Feedback and Criteria from Mean Values, had more modest results. Trial-and-Error Feedback produced a small effect on increasing sensitivity ($d = .19$, $p < .001$) and an even smaller meta-analytic effect on reducing criterion bias ($d = -.10$, $p = .012$), despite this latter effect only being reliable in one analysis (Study 4). Finally, Criteria from Mean Values had a small effect on criterion bias ($d = -.16$, $p < .001$) and no consistent effect on sensitivity ($d = .02$, $p = .575$; see online supplement for full meta-analysis reporting). In all, four interventions could be deemed successful given the rules of the research contest, but the

magnitude and breadth of effectiveness varied substantially across interventions. As a result, researchers or practitioners looking to adapt these interventions for their own purposes may want to start with the most effective interventions when possible.

An Optimistic Interpretation of Results

Results from this research contest offer several reasons for optimism. Most notably, four interventions showed reliable evidence of reducing either decision noise or decision bias in a task known to produce social favoritism in judgment (Axt, Nguyen & Nosek, 2018). The four interventions showed generally consistent effects among both volunteer (Study 1) and paid samples (Studies 2-4), as well as in two different versions of the JBT. These results are encouraging because they suggest that the interventions could be productively scaled up and applied to novel social domains and judgment contexts, and indeed we hope future researchers look to build off this promising evidence and adapt these strategies to reducing other forms of discriminatory behavior.

Moreover, the inclusion of interventions that used similar approaches to the four effective strategies but did *not* impact discrimination provides comparative data that allows for speculation about what may be the "key ingredients" for interventions to change behavior. In particular, past work (Axt & Lai, 2019) has suggested that the level of criterion bias on the JBT is related to one's motivation or ability to ignore socially biasing information (e.g., a face communicating physical attractiveness) while the level of sensitivity on the JBT is more related to one's motivation or ability to process decision-relevant information (e.g., applicant qualifications like GPA). Using this framework, we can then compare across similar interventions that 1) largely changed criterion bias, 2) largely changed sensitivity, 3) changed

both outcomes, or 4) changed neither outcome, to infer what may be the most crucial characteristics of interventions that impacted different components of discrimination.

Features of effective interventions. First, the main characteristic that emerges within the most effective interventions is focusing participant attention on the relevant evaluation criteria. Three of the four interventions that were used across rounds of the contest either provided participants with a strategy or rule for processing applicants' qualifications or asked them to generate a strategy themselves. These results are in line with previous work finding that more effective discrimination-reducing interventions place an emphasis on providing concrete evaluation strategies (Hodson et al., 2002). By setting decision-making standards prior to making judgments, participants in these interventions were perhaps less influenced by social information, like physical attractiveness or political orientation (Quinn, 2020; Milkman et al., 2008; Uhlmann & Cohen, 2005).

Indeed, the two most effective interventions – the Two-out-of-Three Rule and the Criteria Reinforcement Exercise – were the only manipulations that both reduced bias and increased sensitivity within each study, producing particularly large effects on sensitivity (Cohen's $d > 0.9$ in all studies). Broadly, both interventions provided participants with simplified strategies for selecting the best applicants. As a result, the interventions may have derived their effectiveness by simultaneously 1) allowing participants to simplify the decision-making task with a rule that was helpful and simple enough that it could be followed throughout the JBT (Milkman et al., 2009), and 2) diverting attention away from applicant faces and towards applicant qualifications, thereby reducing the impact of physical attractiveness on judgment (Axt & Lai, 2019).

These two interventions also included sample trials that provided feedback, so a potential explanation may be that feedback is the key driving force behind the interventions' success

(Fischoff, 1982). Fortunately, a separate intervention (Trial-and-Error Feedback) included only a practice-and-feedback exercise of 32 trials. This intervention was ultimately deemed successful given the guidelines of the research contest, but in a manner that was quite a bit weaker than the Two-out-of-Three Rule or Criteria Reinforcement Exercise interventions. Specifically, the feedback-only manipulation 1) had a much smaller effect on sensitivity (meta-analytic d = .19 in Studies 1, 2 and 4), and 2) only reliably reduced criterion bias in one of four studies (Study 4), leading to a very weak aggregate effect on criterion bias. Practice and feedback alone may then be slightly helpful for reducing discrimination, as the increase in sensitivity suggests that it allows participants to better process applicants' qualifications. At the same time, the small effect sizes from the intervention indicate that practice and feedback alone did not help participants themselves generate highly effective routes for navigating the JBT, and that merely practicing the task does not inhibit some reliance on social information in judgment (i.e., the intervention still produced a reliable effect of attractiveness-based differences in response criterion in all studies).

Conversely, one intervention – Criteria from Mean Values – only impacted criterion biases and not sensitivity, as the manipulation reduced some component of criterion biases in each study but never consistently changed sensitivity. In this intervention, participants estimated the average value for each of the four JBT qualifications following the encoding phase, and then provided a decision rule to guide their judgments (e.g., "I will only accept candidates if… both GPAs are above 3.5 and the interview score is above 75"). Given the sizable increases in sensitivity that came from the decision rules in the Two-Out-of-Three and Criteria Reinforcement Exercise interventions, the fact that the Criteria from Mean Values intervention did not consistently lead to increased sensitivity suggests either that 1) participants created

effective decision rules, but they were unable to follow them consistently throughout the task, or 2) participants successfully followed their decision rules, but such rules were on average no more effective than performance under control conditions. For example, one participant in this intervention wrote that they would accept applicants if they had one GPA over 3.5 and another GPA over 3.1, as well as at least "good" letters and an interview score above 77. If applied consistently, this rule would have generated overall accuracy (67.2%) that was similar to control conditions (Study 1= 66.5%, Study 2 = 67.8%, Study 4 = 67.0%). Either explanation for the absence of an effect on sensitivity still indicates that participants seem to lack the ability to generate useful or straightforward strategies for making more accurate decisions on the JBT.

However, it is notable that the Criteria from Mean Values intervention still managed to consistently reduce biases in response criterion (meta-analytic d = .16 in Studies 1, 2 and 4), a result that suggests some benefit of merely committing to more specific processing of decision-relevant information. That is, even following a decision rule that is ultimately ineffective at increasing accuracy may still divert attention away from socially biasing information, resulting in a case where there is no change in the amount of unfair treatment given to applicants (i.e., accepting less qualified and rejecting more qualified applicants). Nonetheless, this unfair treatment is more evenly divided among social groups (as evidenced by a reduction in criterion bias). Generating decision rules that neither worsen nor improve accuracy may be particularly helpful in contexts where the correct response is unclear or unknown, an issue we return to below.

Reducing bias, increasing noise. Despite these potential applications, it is worth highlighting how asking participants to attend more to applicant qualifications could produce adverse effects, as two interventions resulted in both lower criterion bias *and* lower sensitivity

(i.e., reduced accuracy). Specifically, the Propositional-Statistical Learning intervention reduced criterion biases and sensitivity in Studies 1-2, and the same pattern emerged in Study 2 for the Single-Criterion Exercise intervention. In the Propositional-Statistical Learning intervention, participants were given a decision rule to accept applicants if 1) Science GPA was at least 3.5 and Interview Scores were at least 75, and 2) if the applicant was "reasonably close" to these cutoffs, accept them if their humanities GPA was over 3.5. Importantly, this strategy should have improved sensitivity. For instance, assuming a Science GPA of 3.3 and Interview Score of 70 were determined as "reasonably close", then following the provided rule should have led to an accuracy rate (70.3%) that was slightly higher than control conditions. That accuracy in this intervention was lowered indicates either 1) the rule was too complex for participants to apply consistently, or 2) the rule was too ambiguous, leading to individual definitions of "reasonably close" that translated into detrimental decision-making strategies.

A similar process may have occurred in the Single-Criterion Exercise intervention, which asked participants to first identify a single criterion they believed was most relevant for their judgment and then generate a decision rule based on that qualification. The intervention had no reliable impact on sensitivity in Study 1 but reduced sensitivity in Study 2. One possible explanation for the Study 2 results is that participants generated rules that would have lessened accuracy. For instance, one participant generated a rule of only admitting applicants with an interview score above 85, which if followed would have decreased accuracy (62.5%) relative to control. Another participant wrote they would reject anyone with a Science GPA below 3.3, which also would have decreased accuracy (54.6%). However, it was possible for the intervention to lead to rules that improved accuracy; for instance, a decision rule that admitted anyone with an interview score of at least 75 would have resulted in 71.9% accuracy. That this

intervention failed to improve sensitivity and even reduced sensitivity in Study 2 suggests that in contexts with more complicated, multi-attribute judgments, many people may lack the ability to effectively simplify the decision-making process in a manner that increases accuracy and may even have a tendency towards generating strategies that hurt overall performance.

Despite reducing accuracy, these two interventions did lessen criterion bias. Again, failing to follow overly difficult rules, or successfully following harmful ones, still seemed to lessen reliance on physical attractiveness in some way during decision-making. Together, the number of interventions that reduced criterion biases but had no impact on (or worsened) sensitivity indicates that there may be a number of ways to divert participants' attention away from biasing social information, but that providing or generating sufficiently simple, effective strategies for navigating the decision-making process is a greater challenge.

Learning from unsuccessful interventions. Finally, interventions that produced no reliable changes in judgment are also informative. For instance, two interventions (Criteria from Min-Max Values and Minimal Threshold Criteria) presented participants with information about applicants' qualifications and asked them to generate thresholds for each qualification that needed to be met in order to provide an "accept" response. Another intervention (Personal Information Avoidance Rule) asked participants to directly indicate what information – either applicant's photos or their qualifications – was most important to consider in order to make an unbiased judgment. Despite a focus on prioritizing applicant qualifications, none of these interventions consistently impacted criterion bias or sensitivity.

Without additional data, it is impossible to confidently identify why each intervention did or did not impact decision bias or noise, but the divergent outcomes that emerged from broadly similar strategies suggests the existence of certain critical features that are needed for

interventions to impact JBT performance. For instance, the only intervention that reduced criterion bias from the "Evaluation Criteria (Self-Determined)" category (without any simultaneous negative consequences, like increasing noise) was the Criteria from Mean Values intervention. Of note, this was also the only intervention that asked participants to actively glean information from the applicant pool during the encoding phase; specifically, participants had to try to learn the mean value of each qualification across the entire applicant pool. This feature was not present in interventions like Criteria from Min-Max Values, which instead simply provided participants with the range of values for each qualification before the encoding phase began (i.e., participants were not asked to learn this information on their own). It is possible that this more active encoding phase trained participants to direct attention away from applicant faces, which in turn lowered criterion bias (though the actual decision rules generated in this intervention may not have been effective at changing sensitivity).

Ultimately, our contest design prioritized researcher freedom in designing interventions, which created significant diversity in approaches and a large amount of comparative data, but this decision came at the expense of greater standardization across intervention strategies. This line of research will thus benefit from future studies that look to isolate the presence or absence of certain intervention components (e.g., active versus passive encoding phases) that may be critical for changing discrimination in social judgment.

A Pessimistic Interpretation of Results

While certain interventions did consistently reduce decision bias and/or decision noise, there remain several reasons to be more pessimistic about our results. For one, the fact that only 4 of 30 interventions were found to impact JBT performance suggests that it is very difficult, even among topic experts, to identify effective routes for addressing discrimination. Our criteria

for eligibility were relatively minimal – interventions could not mention physical attractiveness, could not change the JBT, and had to last less than seven minutes – and participating researchers had full access to the JBT instructions, images, and application profiles. Despite such direct knowledge of our outcome measure, and the freedom to draw from many prior areas of research, only 13% of submitted interventions changed either JBT outcome. This suggests either that 1) researchers were largely unsuccessful in applying strategies from prior studies on discrimination and decision-making to a novel context, 2) researchers successfully recreated these intervention strategies, but the type of discrimination produced by the JBT was particularly resilient to these approaches, or 3) researchers were using the contest study to test novel strategies. Unfortunately, our lack of manipulation checks means we cannot disentangle which null results were due to interventions that did or did not impact the targeted psychological construct, but in either case results still reveal that many common methods for reducing biased judgment could not be applied to a new context.

Another reason for pessimism is the potentially limited application of the interventions that were deemed effective at reducing criterion biases or increasing sensitivity. In particular, each of the four interventions tested in Studies 3-4 had some component that relied on the fact that the JBT has objectively correct and incorrect responses, such as giving feedback on sample trials or providing an effective decision rule that was reverse-engineered from the criteria of qualified versus unqualified applicants. Though decision-makers may strive to make their own evaluation processes as objective as possible and adapt the four interventions identified here, there are many judgments where this is not feasible or even desired.

As a result, the interventions that changed decision noise and/or decision bias may be difficult to apply to decision-making contexts where the objectively correct response is unclear

(e.g., whether to prioritize work experience or educational background in a job applicant) or where applicants cannot be so easily compared across various criteria. Similarly, many judgments involve metrics that may only appear to be objective. For example, two equally glowing reference letters should be viewed much differently if one comes from a supervisor who gives highly positive letters to all trainees versus one who reserves such positivity only for truly exceptional students, though such background information is typically unavailable to decision-makers. In these cases, reducing discrimination may require approaches that are more nuanced than the strategies that were effective on the JBT. Future studies will need to address this concern directly by investigating whether these approaches produce similar results in outcomes where the objectively correct decision is less clear (or non-existent) or where the decision-making criteria are much more difficult to parse.

Aside from what interventions were most likely to change JBT performance, it is also revealing to see what interventions were most likely to be submitted. The two most popular classes of intervention were raising awareness of bias and exposure to counter-stereotypical exemplars. For bias awareness, seven interventions had some component of 1) warning participants about the potential for their judgments to be biased and 2) urging them to make unbiased judgments. None of these manipulations were effective in either reducing criterion bias or increasing sensitivity. These types of awareness interventions have shown some limited success in research on broader judgment biases (e.g., Ludolph & Schulz, 2018), but in general have failed to produce consistent effects (Axt, Casola & Nosek, 2019; Fischhoff, 1977; Jaeger et al., 2020; Lord et al., 1984). Our results suggest that merely raising awareness of bias remains an alluring strategy, even to experts, but one that has yet to show a consistent impact on behavior. Efforts to develop interventions may instead want to prioritize other strategies outside

of raising awareness of bias. Interventions that reduce discrimination may need components that are more concrete than what was typically included in awareness interventions (Wilson & Brekke, 1994). If awareness interventions are pursued further, then some evidence suggests that the form of awareness-raising also ought to be more focused on the form of discrimination being targeted than the interventions in the present studies (Axt et al., 2019).

A potential reason why these more general bias awareness interventions were unsuccessful is that participants either did not believe they would show such biased behavior or became overly confident in their ability to avoid such behavior after reading about it. One intervention – Orchestra Case – is particularly illustrative of this point. Here, participants reviewed a short article on the concept of "automatic discrimination." They then read about an example of how to address this phenomenon, specifically through using blind auditions to ameliorate gender bias in orchestra selections (Goldin & Rouse, 2000). The intervention then invited participants to use their hands as a "screen" to cover the images of applicants' faces as they completed the JBT. If participants followed this instruction, then by definition the intervention would have eliminated criterion biases and should have also increased sensitivity (see Studies 1a-1b in Axt & Lai, 2019). However, the intervention changed neither. That participants eschewed a strategy guaranteeing a reduction in discrimination suggests a certain confidence in one's ability to be an objective evaluator that even direct education about the subtlety of discrimination had a hard time reducing (Pronin et al., 2002). Indeed, participants in all of our control conditions reported high levels of belief that they did not use physical attractiveness in their decision-making (Study 1 = 84.1%, Study 2 = 85.1%, Study 3 = 84.8%, Study 4 = 83.1%), a rate that could represent a near ceiling effect in the capacity to change individual motivations to be unbiased.

Exposure to counter-stereotypical exemplars (e.g., Dasgupta & Greenwald, 2001) was another popular approach, as it also accounted for seven interventions. In these manipulations, participants were asked to associate physically attractive people with either negative actions or incompetence, and to associate physically attractive people with either positive actions or competence (the one exception is the Imagined Contact condition, which only asked participants to imagine interacting with a competent colleague in an effort to downplay reliance on attractiveness when completing the JBT). Again, despite the popularity of the general approach, none of these interventions reduced criterion bias or increased sensitivity, with one intervention (Associative Learning Paradigm) even reducing sensitivity in Study 2.

Using counter-stereotypical exemplars may have been an appealing strategy for researchers because the approach was successful, at least in the short term, in a prior contest study looking to reduce bias on a Black-White IAT (Lai et al., 2014). That similar strategies were not effective for the JBT can be attributed to either a shift in domain (race versus physical attractiveness) or a change in the type of outcome, as the JBT is a measure where responses are under more conscious control relative to the IAT (Axt et al., 2018). These results align with prior findings showing that interventions that reduce one form of intergroup bias (e.g., comparatively automatic forms of prejudice from measures of implicit associations) may not generalize to other forms of intergroup bias (behavioral measures of favoritism; Forscher, Lai, et al., 2019). Progress in reducing discriminatory behavior (or intergroup prejudice) may require greater attention in matching intervention approaches to outcome measures.

Limitations

The biggest limitation of this work is its focus on discrimination operating within only one judgment context (admissions decisions), two social domains (physical attractiveness and

political affiliation), and one outcome measure (the JBT). While these interventions were designed to be applicable to other forms of discrimination (e.g., based on age or religion), whether similar effects emerge in other social dimensions will need to be explored in future work. Similarly, subsequent research in this area will want to investigate whether comparable results emerge in other judgment domains, such as in decisions related to outcomes like promotion, housing, grading, or lending. It would also be valuable to test these interventions in other versions of the JBT. For instance, the current version asked participants to accept approximately 50% of applicants. Prior work on the relationship between prejudice and threat suggests that biases may be exacerbated under conditions of scarcity (e.g., Krosch & Amodio, 2017; Rodenheffer et al., 2012), meaning the effects of these manipulations may look different when acceptances are rarer. In short, the generalizability of this work to other forms of discrimination, other types of judgment, or even other versions of the JBT is currently unclear, and we hope that insights from this project will be tested in such contexts.

It is also possible that the interventions identified here as reducing decision bias and/or decision noise are not easily adapted to other types of discriminatory behavior, such as outcomes that lack an objectively correct response (e.g., decisions on whether to pursue friendships with outgroup members), outcomes that have a correct response but are completed under intense time pressure (e.g., the First Person Shooter Task; Correll et al., 2002), or outcomes that are under less conscious control (e.g., biases in nonverbal behavior; LaCosse & Plant, 2020). Conversely, it is also possible that some of the interventions identified as ineffective in the present work may show success in other forms of discrimination; for example, the association training and counter-stereotypic exemplar interventions could reduce intergroup biases in impression formation (e.g., Branscombe & Smith, 1990). Moreover, our criteria for

intervention eligibility (e.g., lasting less than seven minutes, not allowing changes to the JBT) precluded several prominent bias-reducing strategies from being tested, such as joint evaluation (Bohnet et al., 2016) or allowing participants to first place candidates in a "shortlist" that could be revisited later (e.g., Lucas et al., 2021). In addition, the lack of manipulation checks in these studies may impede future work on this topic, as it is possible that some interventions would have changed discriminatory behavior if the targeted psychological construct had been effectively altered. However, our design makes it impossible to differentiate between interventions that simply failed to change the targeted construct from those that successfully manipulated the construct of interest, but this change simply did not translate into differences in JBT performance.

The present work also cannot speak to the duration of intervention effectiveness. Follow-up findings from the prior research contest (Lai et al., 2016) found no evidence that interventions that changed performance on an IAT administered immediately after the manipulation had any effect on an IAT completed twenty-four hours later, and the same may be true of the interventions tested here. If these interventions have only short-term effects, it would limit the potential application of our results. Many instances of discrimination unfold over longer periods of time (e.g., reviewing applications for a job opening over several weeks), meaning that any training or interventions would need to persist throughout the evaluation process. Given that responses on the JBT are more controlled than those on the IAT, there is some reason for optimism about the possibility that interventions can retain their effectiveness over time; for example, it may not be overly challenging for participants to remember and apply the Two-out-of-Three Rule over JBT sessions lasting several days or weeks, particularly if they

were provided with a brief reminder of the rule any time they viewed applicants. However, testing this assumption directly should be a focus of future studies.

Finally, the judgments in the JBT were all hypothetical, and were made in a context where participants knew they were completing a study. Though Study 4 results lend support to the idea that even participants with self-reported hiring experience demonstrate these biases, it is an open and important question to know whether biases of comparable magnitude exist within real-world judgments that have actual consequences (e.g., Rooth, 2009) and to what extent these interventions are similarly effective in field settings. We hope this work, which was completed under lab conditions that maximized internal validity, will help guide future efforts seeking to reduce discriminatory behavior in real-world judgments. See Table 6 for more details about the limitations of this work.

Conclusion

In the largest comparative study to date of interventions to reduce discrimination, only four of 30 reduced decision bias, reduced decision noise, or both. These interventions mostly centered on focusing participant attention on decision-relevant criteria, though the divergent outcomes that emerged from largely similar approaches suggests that effectively changing discriminatory behavior may depend on specific characteristics (e.g., providing a decision rule that is both effective and easy to remember). Popular approaches from the intergroup relations or decision-making literature, such as invoking accountability, raising awareness, and viewing counter-stereotypic exemplars, failed to impact our measure of discrimination. We hope that the present work, along with our open materials and data (https://osf.io/wk2s9/) can spur future investigations that better identify how and why certain interventions did or did not impact

Table 6.

Assessment of Limitations.

Dimensions	Assessment
	Internal Validity
Is the phenomenon diagnosed with experimental methods?	Yes
Is the phenomenon diagnosed with longitudinal methods?	No
Were the manipulations validated with manipulation checks, pretest data, or outcome data?	Manipulation of applicant attractiveness in the JBT was validated in a prior paper using the JBT (Axt, Nguyen & Nosek, 2018). Manipulations of intervention strategies were all based in past research but not directly validated prior to inclusion in the research contest. Lastly, there was no manipulation checks to ensure that the participants were following the interventions' instructions.
What possible artifacts were ruled out?	Study 2 rules out the potential artifact that interventions are only effective on a volunteer sample. Study 3 rules out the potential artifact that interventions can only show effectiveness in a context with a single piece of biasing social information. Study 4 rules out the potential artifact that interventions are not effective for trained decision-makers (i.e., those with self-reported hiring experience).
	Statistical Validity
Was the statistical power at least 80%?	Yes, for all analyses.
Was the reliability of the dependent measure established in this publication or elsewhere in the literature?	Reliability of the dependent measure -- the JBT -- has been shown in a prior paper introducing the measure (Axt, Nguyen & Nosek, 2018).
If covariates are used, have the researchers ensured they are not affected by the experimental manipulation before including them in comparisons across experimental groups?	Not applicable.

Were the distributional properties of the variables examined and did the variables have sufficient variability to verify effects?	Yes
	Generalizability to Different Methods
Were different experimental manipulations used?	Each intervention only had one version, though each of our seven broader intervention strategies contained at least two interventions. We also used two versions of the JBT - one using only physical attractiveness and one using both physical attractiveness and political affiliation.
	Generalizability to Field Settings
Was the phenomenon assessed in a field setting?	No
Are the methods artificial?	Yes, the methods are highly artificial.
	Generalizability to Times and Populations
Are the results generalizable to different years and historic periods?	This was not tested, but, given changing contexts of social biases, results may be different for other historic periods.
Are the results generalizable across populations (e.g., different ages, cultures, or nationalities)?	This was not tested, except for varying sample sources (volunteer versus paid) across studies. In Study 4, we also tested the effects with a sample of those with self-reported hiring experience as opposed to lay participants in other studies. Given that all studies included US samples, results will likely differ in other populations.
	Theoretical Limitations
What are the main theoretical limitations?	The main theoretical limitation is the lack of investigation of the underlying processes leading to our results. The current data are suggestive but not conclusive regarding what psychological mechanisms are responsible for changing performance on the JBT.

discriminatory judgment, and apply findings to new forms of discrimination. In all, results suggest that even for experts, reducing discrimination remains a challenge, and that greater attention may need to be devoted to the development and validation of effective interventions before such methods are applied outside of the laboratory.

References

Ameri, M., Schur, L., Adya, M., Bentley, S., McKay, P., & Kruse, D. (2015). *The Disability Employment Puzzle: A Field Experiment on Employer Hiring Behavior*. https://doi.org/10.3386/w21560

Axelrod, R. (1980). Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution*, *24*(1), 3–25. https://doi.org/10.1177/002200278002400101

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. https://doi.org/10.1177/0146167218814003

Axt, J. R., & Johnson, D. J. (2021). Understanding mechanisms behind discrimination using diffusion decision modeling. *Journal of Experimental Social Psychology*, *95*, 104134. https://doi.org/10.1016/j.jesp.2021.104134

Axt, J. R., & Lai, C. K. (2019). Reducing discrimination: A bias versus noise perspective. *Journal of Personality and Social Psychology*, *117*(1), 26–49. https://doi.org/10.1037/pspa0000153

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

Axt, J. R., Yang, J., & Deshpande, H. (2022). Misplaced intuitions in interventions to reduce attractiveness-based discrimination. *Personality and Social Psychology Bulletin*, 01461672221074748. https://doi.org/10.1177/01461672221074748

Berggren, N., Jordahl, H., & Poutvaara, P. (2010). The looks of a winner: Beauty and electoral success. *Journal of public economics*, *94*(1-2), 8-15.

Bieleke, M., Keller, L., & Gollwitzer, P. M. (2021). If-then planning. *European Review of Social Psychology*, *32*(1), 88–122. https://doi.org/10.1080/10463283.2020.1808936

Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology*, *72*(3), 544–557. https://doi.org/10.1037/0022-3514.72.3.544

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66*(1), 5–20. https://doi.org/10.1037/0022-3514.66.1.5

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. Journal of Personality and Social *Psychology, 81*(5), 828–841. https://doi.org/10.1037/0022-3514.81.5.828

Bohnet, I., Van Geen, A., & Bazerman, M. (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, *62*(5), 1225-1234.

Branscombe, N. R., & Smith, E. R. (1990). Gender and racial stereotypes in impression formation and social decision-making processes. *Sex Roles*, *22*(9), 627–647. https://doi.org/10.1007/BF00288239

Bruneau, E., Kteily, N., & Falk, E. (2018). Interventions highlighting hypocrisy reduce collective blame of Muslims for individual acts of violence and assuage anti-Muslim hostility. *Personality and Social Psychology Bulletin*, *44*(3), 430–448. https://doi.org/10.1177/0146167217744197

Bruneau, E. G., Kteily, N. S., & Urbiola, A. (2020). A collective blame hypocrisy intervention enduringly reduces hostility towards Muslims. *Nature Human Behaviour*, *4*(1), Article 1. https://doi.org/10.1038/s41562-019-0747-7

Byrne, D. E. (1971). *The attraction paradigm*. Academic Press.

Capers, Q., McDougle, L., & Clinchot, D. M. (2018). Strategies for achieving diversity through medical school admissions. *Journal of Health Care for the Poor and Underserved*, *29*(1), 9–18. https://doi.org/10.1353/hpu.2018.0002

Chang, E. H., Milkman, K. L., Gromet, D. M., Rebele, R. W., Massey, C., Duckworth, A. L., & Grant, A. M. (2019). The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences*, *116*(16), 7778–7783. https://doi.org/10.1073/pnas.1816076116

Chua, K.-W., & Freeman, J. B. (2021). Facial stereotype bias is mitigated by training. *Social Psychological and Personality Science, 12*(7), 1335–1344. https://doi.org/10.1177/1948550620972550

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*(6), 1314–1329.

Crisp, R. J., & Hewstone, M. (2007). Multiple social categorization. *Advances in experimental social psychology, Vol 39* (pp. 163–254). Elsevier Academic Press. https://doi.org/10.1016/S0065-2601(06)39004-1

Crisp, R. J., Stathi, S., Turner, R. N., & Husnu, S. (2009). Imagined intergroup contact: Theory, paradigm and practice. *Social and Personality Psychology Compass*, *3*(1), 1–18. https://doi.org/10.1111/j.1751-9004.2008.00155.x

Crisp, R. J., & Turner, R. N. (2012). The Imagined Contact Hypothesis. *Advances in Experimental Social Psychology, 46*, 125–182. https://doi.org/10.1016/B978-0-12-394281-4.00003-9

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*(5), 800–814. https://doi.org/10.1037/0022-3514.81.5.800

Dobbin, F., & Kalev, A. (2016). Why diversity programs fail. *Harvard Business Review*, *94*(7).

Doleac, J. L., & Stein, L. C. D. (2013). The Visible Hand: Race and online market outcomes. *The Economic Journal*, *123*(572), F469–F492. https://doi.org/10.1111/ecoj.12082

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*(4), 315–319. https://doi.org/10.1111/1467-9280.00262

Dovidio, J. F., & Gaertner, S. L. (2010). Intergroup bias. In *Handbook of social psychology, Vol. 2, 5th ed* (pp. 1084–1121). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470561119.socpsy002029

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, *9*(2), 1–22. https://doi.org/10.1257/app.20160213

Fath, S., Larrick, R. P., & Soll, J. B. (2022). Blinding curiosity: Exploring preferences for "blinding" one's own judgment. *Organizational Behavior and Human Decision Processes*, *170*, 104135. https://doi.org/10.1016/j.obhdp.2022.104135

Feinberg, M., & Willer, R. (2019). Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, *13*(12), e12501. https://doi.org/10.1111/spc3.12501

Feingold, A. (1992). *Good-looking people are not what we think.* https://doi.org/10.1037/0033-2909.111.2.304

Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on ratemyprofessors. com. *Assessment & Evaluation in Higher Education*, *33*(1), 45-61.

Feng, Z., Liu, Y., Wang, Z., & Savani, K. (2020). Let's choose one of each: Using the partition dependence effect to increase diversity in organizations. *Organizational Behavior and Human Decision Processes*, *158*, 11-26.

Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and how implicit first impressions can be updated. *Current Directions in Psychological Science*, *28*(4), 331–336. https://doi.org/10.1177/0963721419835206

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(2), 349-358.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slavic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422-444). Cambridge, England: Cambridge Univ. Press.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fisher, P., Risavy, S., Robie, C., Konig, C., Christiansen, N., Tett, R., & Simonet, D. (2020). Selection myths: A conceptual replication of HR professionals' beliefs about effective human resource practices in the United States and Canada. *Journal of Personnel Psychology*, *20*, 51–60.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, *90*(4), 715–741. https://doi.org/10.1257/aer.90.4.715

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Chapter Two - Moral Foundations Theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Academic Press. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037/0033-295x.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

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

Grigoryan, L., Cohrs, J. C., Boehnke, K., van de Vijver, F. A. J. R., & Easterbrook, M. J. (2022). Multiple categorization and intergroup bias: Examining the generalizability of three theories of intergroup relations. *Journal of Personality and Social Psychology, 122*(1), 34–52. https://doi.org/10.1037/pspi0000342

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion* (1st ed). Pantheon Books.

Hales, A. H. (2023). One-tailed tests: Let's do this (responsibly). *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000610

Hameiri, B., & Moore-Berg, S. L. (2022). Intervention tournaments: An overview of concept, design, and implementation. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 17456916211058090. https://doi.org/10.1177/17456916211058090

Hamermesh, D. S., & Biddle, J. (1993). Beauty and the labor market. *American Economic Review: A Journal of the American Economic Association. 84*(5), 1174-1194.

He, J. C., Kang, S. K., & Lacetera, N. (2021). Opt-out choice framing attenuates gender differences in the decision to compete in the laboratory and in the field. *Proceedings of the National Academy of Sciences*, *118*(42), e2108337118.

Hester, N., & Gray, K. (2018). For Black men, being tall increases threat stereotyping and police stops. *Proceedings of the National Academy of Sciences*, *115*, 2711-2715.

Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin*, *28*(4), 460–471. https://doi.org/10.1177/0146167202287004

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. https://doi.org/10.1111/spc3.12111

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019). Explaining the persistent influence of facial cues in social decision-making. *Journal of Experimental Psychology: General*, *148*(6), 1008–1021. https://doi.org/10.1037/xge0000591

Jaeger, B., Todorov, A. T., Evans, A. M., & van Beest, I. (2020). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology*, *90*, 104004. https://doi.org/10.1016/j.jesp.2020.104004

Johnson, S. K., & Chan, E. (2019). Can looks deceive you? Attractive decoys mitigate beauty is beastly bias against women. *Archives of Scientific Psychology*, *7*(1), 60. https://doi.org/10.1037/arc0000066

Jones, R., et al. (2022). Proper vs. Improper Jury Instructions, Guide Fed. Civ. Trials & Ev. Ch. 15-B.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux; US.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kopelman, S. (2020). Tit for Tat and beyond: The legendary work of Anatol Rapoport. *Negotiation and Conflict Management Research*, *13*(1), 60–84. https://doi.org/10.1111/ncmr.12172

LaCosse, J., & Plant, E. A. (2020). Internal motivation to respond without prejudice fosters respectful responses in interracial interactions. *Journal of Personality and Social Psychology*, *119*(5), 1037–1056. https://doi.org/10.1037/pspi0000219

Lai, C.K. & Lisnek, J.A. (2023). The impact of implicit bias-oriented diversity training on police officers' beliefs, motivations, and actions. *Psychological Science*. Advance online publication.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., … Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. https://doi.org/10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M., Burns, M., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time.

Journal of Experimental Psychology: General, *145*(8), 1001–1016. https://doi.org/10.1037/xge0000179

Langton, S. R. H., Law, A. S., Burton, A. M., & Schweinberger, S. R. (2008). Attention capture by faces. *Cognition*, *107*(1), 330–342. https://doi.org/10.1016/j.cognition.2007.07.012

Lehman, D. R., & Nisbett, R. E. (1990). A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology*, *26*, 952-960.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255

Lerner, J. S., & Tetlock, P. E. (2003). Bridging individual, interpersonal, and institutional approaches to judgment and choice: The impact of accountability on cognitive bias. In S. Schneider & J. Shanteau (Ed.), *Emerging perspectives on judgment and decision research* (pp. 431-457) . Cambridge, Cambridge University Press.

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, 26(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Lippens, L., Vermeiren, S., & Baert, S. (2023). The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments. *European Economic Review, 151*, 104315.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231–1243.

Lucas, B. J., Berry, Z., Giurge, L. M., & Chugh, D. (2021). A longer shortlist increases the consideration of female candidates in male-dominant domains. *Nature Human Behaviour, 5*(6), 736-742.

Lu, C., & Wan, C. (2018). Cultural self-awareness as awareness of culture's influence on the self: Implications for cultural identification and well-being. *Personality & Social Psychology Bulletin*, *44*(6), 823–837. https://doi.org/10.1177/0146167217752117

Lu, C., Lee, I.-C., & Wan, C. (2023). Intergroup benefits of metacognitive cultural self? Cultural self-awareness and multicultural involvement on attitudes towards migrants. *Group Processes & Intergroup Relations*, 136843022211475. https://doi.org/10.1177/13684302221147509

Ludolph, R., & Schulz, P. J. (2018). Debiasing health-related judgments and decision making: A systematic review. *Medical Decision Making*, *38*(1), 3–13. https://doi.org/10.1177/0272989X17716672

Lueke, A., & Gibson, B. (2015). Mindfulness meditation reduces implicit age and race bias: The role of reduced automaticity of responding. *Social Psychological and Personality Science*, *6*(3), 284–291. https://doi.org/10.1177/1948550614559651

Lueke, A., & Gibson, B. (2016). Brief mindfulness meditation reduces discrimination. *Psychology of Consciousness: Theory, Research, and Practice*, *3*(1), 34–44. https://doi.org/10.1037/cns0000081

Lutz, A., Jha, A. P., Dunne, J. D., & Saron, C. D. (2015). Investigating the phenomenological matrix of mindfulness-related practices from a neurocognitive perspective. *American Psychologist*, *70*(7), 632–658. https://doi.org/10.1037/a0039585

Marini, M., Rubichi, S., Sartori, G. (2012). The role of self-involvement in shifting IAT effects. *Experimental Psychology*, *59*(6), 348-354, https://doi.org/10.1027/1618-3169/a000163

Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality*

and Social Psychology Bulletin, *36*(4), 512–523. https://doi.org/10.1177/0146167210362789

Miles, A., Charron-Chénier, R., & Schleifer, C. (2019). Measuring automatic cognition: Advancing dual-process research in sociology. *American Sociological Review*, *84*(2), 308–333. https://doi.org/10.1177/0003122419832497

Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination: An audit study in academia. *Psychological Science*, *23*(7), 710–717. https://doi.org/10.1177/0956797611434539

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science*, *4*(4), 379–383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

Milkman, K. L., Rogers, T., & Bazerman, M. H. (2008). Harnessing our inner angels and demons: What we have learned about want/should conflicts and how that knowledge can help us reduce short-sighted decision making. *Perspectives on Psychological Science*, *3*(4), 324–338. https://doi.org/10.1111/j.1745-6924.2008.00083.x

Mobius, M. M., & Rosenblat, T. S. (2006). Why beauty matters. *American Economic Review*, *96*(1), 222-235.

Monk Jr, E. P., Esposito, M. H., & Lee, H. (2021). Beholding inequality: Race, gender, and returns to physical attractiveness in the United States. *American Journal of Sociology*, *127*(1), 194-241.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109

Nault, K. A., Pitesa, M., & Thau, S. (2020). The attractiveness advantage at work: A cross-disciplinary integrative review. *Academy of Management Annals*, *14*(2), 1103-1139.

Norton, M. I., Sommers, S. R., Apfelbaum, E. P., Pura, N., & Ariely, D. (2006). Color blindness and interracial interaction: Playing the political correctness game. *Psychological Science*, *17*(11), 949–953. https://doi.org/10.1111/j.1467-9280.2006.01810.x

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology*, *87*(6), 817–831. https://doi.org/10.1037/0022-3514.87.6.817

Oettingen, G. (2015). *Rethinking Positive Thinking: Inside the New Science of Motivation*. Penguin Publishing Group.

Olson, I. R., & Marshuetz, C. (2005). Facial attractiveness is appraised in a glance. *Emotion*, *5*(4), 498–502. https://doi.org/10.1037/1528-3542.5.4.498

Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology, 108*(5), 937–975.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*(1), 339–367. https://doi.org/10.1146/annurev.psych.60.110707.163607

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology*, *72*(1), 533–560. https://doi.org/10.1146/annurev-psych-071620-030619

Pope, D. G., Price, J., & Wolfers, J. (2018). Awareness reduces racial bias. *Management Science*, *64*(11), 4988–4995. https://doi.org/10.1287/mnsc.2017.2901

Prentice, M., Jayawickreme, E., Hawkins, A., Hartley, A., Furr, R. M., & Fleeson, W. (2019). Morality as a basic psychological need. *Social Psychological and Personality Science*, *10*(4), 449–460. https://doi.org/10.1177/1948550618772011

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *43*(4), 565–578. https://doi.org/10.1016/j.jesp.2006.05.011

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381. https://doi.org/10.1177/0146167202286008

Quadlin, N. (2018). *The Mark of a Woman's Record: Gender and Academic Performance in Hiring*. https://doi.org/10.1177/0003122418762291

Quinn, D. M. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis*, *42*(3), 375–392. https://doi.org/10.3102/0162373720932188

Ritchie, K., Palermo, R., & Rhodes, G. (2017). Forming impressions of facial attractiveness is mandatory. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-00526-9

Rivera, L. A. (2015). Go with your gut: Emotion and evaluation in job interviews. *American Journal of Sociology*, *120*(5), 1339–1389. https://doi.org/10.1086/681214

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59. https://doi.org/10.1007/BF00055564

Sassenberg, K., & Moskowitz, G. B. (2005). Don't stereotype, think different! Overcoming automatic stereotype activation by mindset priming. *Journal of Experimental Social Psychology*, *41*(5), 506–514. https://doi.org/10.1016/j.jesp.2004.10.002

Scopelliti, I., Morewedge, C. K., McCormick, E., Min, H. L., Lebrecht, S., & Kassam, K. S. (2015). Bias Blind Spot: Structure, measurement, and consequences. *Management Science*, *61*(10), 2468–2486. https://doi.org/10.1287/mnsc.2014.2096

Shen, X., & Ferguson, M. J. (2021). How resistant are implicit impressions of facial trustworthiness? When new evidence leads to durable updating. *Journal of Experimental Social Psychology*, *97*, 104219. https://doi.org/10.1016/j.jesp.2021.104219

Shen, X., Mann, T. C., & Ferguson, M. J. (2020). Beware a dishonest face?: Updating face-based implicit impressions using diagnostic behavioral information. *Journal of Experimental Social Psychology*, *86*. https://doi.org/10.1016/j.jesp.2019.103888

Skitka, L. J., Hanson, B. E., Morgan, G. S., & Wisneski, D. C. (2021). The psychology of moral conviction. *Annual Review of Psychology*, *72*, 347–366. https://doi.org/10.1146/annurev-psych-063020-030612

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review*, *11*(1), 22–40. https://doi.org/10.5465/amr.1986.4282615

Soll, J.B., Milkman, K.L. and Payne, J.W. (2015). A User's Guide to Debiasing. In *The Wiley Blackwell Handbook of Judgment and Decision Making* (eds G. Keren and G. Wu). https://doi.org/10.1002/9781118468333.ch33Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology*, *56*(4), 283–294. https://doi.org/10.1027/1618-3169.56.4.283

Thaler, R. H., & Benartzi, S. (2004). Save More Tomorrow™: Using behavioral economics to increase employee saving. *Journal of Political Economy*, *112*(S1), S164–S187. https://doi.org/10.1086/380085

Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition*, *13*(6), 657–665. https://doi.org/10.1080/13506280500410949

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, *16*(6), 474–480. https://doi.org/10.1111/j.0956-7976.2005.01559.x

Vezzali, L., Capozza, D., Stathi, S., & Giovannini, D. (2012). Increasing outgroup trust, reduced infrahumanization, and enhancing future contact intentions via imagined intergroup contact. *Journal of Experimental Social Psychology*, *48*. (1), 437-440. https://doi.org/10.1016/j.jesp.2011.09.008

Vezzali, L., Stathi, S., Crisp, R. J., Giovannini, D., Capozza, D., & Gaertner, S. L. (2015). Imagined intergroup contact and common ingroup identity: An integrative approach. *Social Psychology*, *46*(5), 265–276. https://doi.org/10.1027/1864-9335/a000242

Voelkel, J. G., Mernyk, J., & Willer, R. (2021). Navigating the Progressive Paradox: The Effects of Value Reframing on Support for Economically Progressive Candidates. *SocArXiv*. https://doi.org/10.31235/osf.io/kdjc5

Voelkel, J., Stagnaro, M., Chu, J., Pink, S., Mernyk, J., Redekopp, C., ... & Willer, R. (2022). Megastudy identifying successful interventions to strengthen Americans' democratic attitudes. Northwestern University: Evanston, IL, USA.

Webster, D. M., Richter, L., & Kruglanski, A. W. (1996). On leaping to conclusions when feeling tired: Mental fatigue effects on impressional primacy. *Journal of Experimental Social Psychology*, *32*(2), 181–195. https://doi.org/10.1006/jesp.1996.0009

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117–142. https://doi.org/10.1037/0033-2909.116.1.117

Wissler, R. L., Kuehn, P. F., & Saks, M. J. (2000). Instructing jurors on general damages in personal injury cases: Problems and possibilities. *Psychology, Public Policy, and Law*, *6*(3), 712-742.