Empirical Research Paper

Reducing Social Judgment Biases May Require Identifying the Potential Source of Bias

Personality and Social Psychology Bulletin
1–20

DOI: 10.1177/0146167218814003
journals.sagepub.com/home/pspb
SAGE

Jordan R. Axt[1,2], Grace Casola[3], and Brian A. Nosek[3,4]

Abstract

Social judgment is shaped by multiple biases operating simultaneously, but most bias-reduction interventions target only a single social category. In seven preregistered studies (total $N > 7{,}000$), we investigated whether asking participants to avoid one social bias affected that and other social biases. Participants selected honor society applicants based on academic credentials. Applicants also differed on social categories irrelevant for selection: attractiveness and ingroup status. Participants asked to avoid potential bias in one social category showed small but reliable reductions in bias for that category ($r = .095$), but showed near-zero bias reduction on the unmentioned social category ($r = .006$). Asking participants to avoid many possible social biases or alerting them to bias without specifically identifying a category did not consistently reduce bias. The effectiveness of interventions for reducing social biases may be highly specific, perhaps even contingent on explicitly and narrowly identifying the potential source of bias.

Keywords

bias, social judgment, discrimination, prejudice

Received December 18, 2017; revision accepted October 12, 2018

Intentional or not, people use ostensibly irrelevant social information to evaluate others. Social judgment biases may arise from reliance on physical features, such as when overweight candidates are less likely to be selected for a position than nonoverweight candidates (Pingitore, Dugoni, Tindale, & Spring, 1994). Biases may also emerge when people use group identity to inform evaluation, such as when foreign applicants are rated as less hirable than native applicants (Hosoda & Stone-Romero, 2010).

There are many interventions that could reduce such biases (e.g., Gollwitzer, 1999; Lerner & Tetlock, 1999). One of the most direct ways is to alert people to a potential bias in judgment, so they can adjust their decision making to avoid it. Models of biased judgment highlight awareness of the source of bias as necessary for reducing biased behavior (Wegener & Petty, 1997; Wilson & Brekke, 1994). Interventions that alert people to biases in their judgment have worked in some cases (Golding, Fowler, Long, & Latta, 1990; Schul, 1993) but not others (Wegner, Coulton, & Wenzlaff, 1985; Wetzel, Wilson, & Kort, 1981). In the present research, we use a paradigm where asking participants to avoid a specific bias in judgment reliably reduces social judgment biases (Axt & Nosek, 2018).

Most research on social judgment biases examines a single category at a time, such as race, age, or weight. But, in reality, people have identities on all those things at once, and social judgment may be influenced by an interplay of evaluations on a variety of social categories. An obvious question is whether interventions are effective at reducing bias across multiple social categories simultaneously, or are specific to individual categories, leaving other social biases unchanged. The answer has practical and theoretical implications. If an intervention is effective for only a single category, then application of that intervention will have limited scope.

There are different theoretical implications if bias-reduction interventions affect single or multiple categories. If bias reduction is limited to a single category, it implies that attention and correction processes are responding to an identified social category. If bias reduction occurs across multiple categories, it implies that decision processes are attending to the relevant information for judgment and "putting aside"

[1]Duke University, Durham, NC, USA
[2]Project Implicit, Seattle, WA, USA
[3]University of Virginia, Charlottesville, USA
[4]Center for Open Science, Charlottesville, VA, USA

Corresponding Author:
Jordan R. Axt, Duke University, 334 Blackwell St., Suite 320, Durham, 27701, NC, USA.
Email: jordan.axt@duke.edu

information from other irrelevant categories. In the present research, we investigated whether an intervention asking people to avoid showing bias had a constrained impact on the social categories identified in the intervention itself or a general impact on social categories whether or not they were explicitly identified in the intervention.

Existing Evidence

Research on intersectionality has examined how multiple social categories independently or interactively influence evaluation (Cole, 2009; Kang & Bodenhausen, 2015). For example, an analysis of over 500,000 employees revealed that those belonging to multiple minority groups (e.g., having both a disability and belonging to a racial minority) received lower pay than employees belonging to a single minority group (Woodhams, Lupton, & Cowling, 2015). Similar results emerged in an analysis of payment toward employees who were both racial and gender minorities (Greenman & Xie, 2008), and a field experiment found that applicants with multiple stigmatized identities were rated as less employable than applicants with a single threat-related identity (Derous, Ryan, & Serlie, 2015).

These examples show a healthy literature concerning the presence of multiple social biases toward single targets. However, we have not found literature investigating how interventions to reduce such biases influence single or multiple social categories simultaneously. The closest are studies investigating malleability of implicit attitudes (Joy-Gaba & Nosek, 2010; Lai et al., 2014). For example, in Lai et al. (2014), priming the concept of multiculturalism was moderately effective at reducing implicit preferences for White versus Black people, but did not alter implicit preferences for White versus Hispanic people or White versus Asian people. To expand this literature, we investigated whether interventions to reduce social judgment biases had effectiveness limited to the social category explicitly identified in the intervention or whether they reduced bias in general toward the targets of judgment.

Theoretical Expectations

Wilson and Brekke's (1994) "mental contamination" model identifies the origins of biased judgment and processes necessary to counteract the expression of bias. Mental contamination occurs when behaviors are influenced by factors that exist outside of conscious intention or awareness. The model presumes four necessary features to avoid mental contamination: (a) awareness of unwanted processing related to the judgment, (b) motivation to correct bias, (c) awareness of direction and magnitude of bias, and (d) ability to adjust responses. Failure to meet any condition results in biased judgment. If decision makers are not alerted to the unwanted bias and able to adjust their responses effectively, then there is no opportunity for motivation and corrective processes to engage.

For situations in which two social categories simultaneously contaminate evaluation of individual targets, the model would most directly predict that alerting people to a bias toward one social category could reduce bias toward that category but not another social category. However, Wilson and Brekke argue that awareness of bias need not come from an external source. For example, alerting people to a bias concerning one social category might lead people to spontaneously notice other potential biases concerning other social categories. Moreover, the model is not definitive on specificity of awareness. For instance, for reducing political orientation bias, would asking people to avoid showing a bias toward "people from different categories" be as effective as asking to people to avoid showing a bias toward "people from different political parties"? And, would warning about another category be sufficient to initiate corrective processes that would reduce political biases too? The model does not directly anticipate or exclude such possibilities, but empirical evidence would help refine the theory.

A second theoretical perspective focuses on activation of motivational processes to reduce reliance on stereotypes and potentially decrease biased behavior (Moskowitz, Gollwitzer, Wasel, & Schaal, 1999). Similar to the Wilson and Brekke (1994) model, this perspective argues that activation of the social category is necessary, but it may not require "awareness" of bias in the conventional sense. For example, one study suggests that activating a social category by brief presentation of an outgroup face is sufficient to invoke correction (Moskowitz, Salomon, & Taylor, 2000). The emphasis of this motivational account is that the activation or awareness of potential bias initiates motivated corrective processes to avoid expression of bias. For example, White participants who thought of a time they failed to act in a racially egalitarian manner later showed increased inhibition of racial stereotypes (Moskowitz & Li, 2011), and White participants told they held racial biases in implicit evaluations displayed increased inhibition when processing race-related information (Monteith, Ashburn-Nardo, Voils, & Czopp, 2002). Relatedly, interventions warning of the possibility of racial bias increased concern about racial discrimination (Devine, Forscher, Austin, & Cox, 2012; Forscher, Mitamura, Dix, Cox, & Devine, 2017).

These motivational accounts are also nonspecific as to whether alerting people to one bias would create reductions in other simultaneous biases. On one hand, the empirical studies virtually always activate the same social category that is assessed for change in behavior. This implies that activation of the motivation to be unbiased is restricted to the activated social category. Yet, the theoretical models leave open the possibility of more general impact of motivational processes. The interventions may highlight a particular social category of potential bias but activate general motivation to be unbiased, leading to corrective processes that reduce

reliance on all irrelevant social information in judgment. Some preliminary support for this possibility comes from findings that priming a general mind-set to "think different" was sufficient for reducing stereotype activation on a lexical decision task (Sassenberg & Moskowitz, 2005).

The reviewed models of bias correction have different emphases but share an expectation that awareness of potential bias (coupled with an ability and motivation to address it) is important for changing biased behavior, and they share a lack of specificity for whether interventions targeting biases for one social category are sufficient to reduce bias toward another social category simultaneously. As such, both models will be improved by evaluating whether bias-reduction interventions must alert people to the particular social category to reduce bias.

The Present Work

In all studies, participants completed a Judgment Bias Task (JBT; Axt, Nguyen, & Nosek, 2018). In the JBT, participants evaluate profiles for an outcome, here admission to an honor society. Each profile includes quantified criteria relevant for evaluation (grade point average [GPA], interview scores) and social information that is ostensibly irrelevant. Criteria are to be weighed equally in judgments, and profiles are constructed so that some are more qualified than others. Participants are assessed on their ability to identify the more over the less qualified applicants and their criterion (c) for accepting an applicant to the honor society using signal detection analysis.

The JBT assesses how social information affects evaluation by comparing the criterion value for profiles belonging to each social category. A lower criterion value means more leniency—specifically, a greater proportion of errors falsely admitting less qualified applicants relative to errors falsely rejecting more qualified applicants. In the academic JBT, a lower criterion means both qualified and unqualified applicants from a social group are more likely to be admitted to the honor society, and bias is evident when criterion is lower for applicants from certain social groups over others.

In these studies, applicants were presented with two forms of social information known to create favoritism: faces varying in physical attractiveness (Feingold, 1992) and an image indicating ingroup or outgroup status (Mullen, Brown, & Smith, 1992). Each applicant was evaluated in the presence of both forms of social information, and the influence of each bias could be assessed independently due to a fully crossed design.

In Studies 1a and 1b, some participants received an intervention that mentioned the potential for ingroup favoritism in judgment and asked them to avoid displaying such favoritism. We investigated whether that intervention reduced favoritism for ingroup members and favoritism for more physically attractive people. In Studies 2a and 2b, some participants received an intervention that asked them to avoid displaying either one or both of the physical attractiveness or ingroup biases. We then investigated whether the bias warning interventions reduced bias for the specific category only or whether it also reduced bias for the unmentioned category. The evidence indicates that the interventions are effective at reducing biases for the identified category but not for others. In Study 3, we tested warning about multiple potential biases (e.g., race, sexual orientation) in addition to biases concerning physical attractiveness and ingroup membership mirroring standard language in Equal Employment Opportunity Commission (EEOC) policies. We found that providing a long list of potential biases was less effective than only warning about the attractiveness and ingroup biases. Finally, in Study 4, we tested whether warning about a potential bias for social categories in general would reduce bias toward all categories because of its breadth or toward no categories because of its lack of specificity. The evidence suggests the latter.

Studies 1a and 1b

In Study 1a, profiles were presented with more or less physically attractive faces and images indicating applicants came from one's own or a rival university. Study 1b used the same faces, but replaced university with political affiliation. In both studies, participants in control conditions completed the JBT without additional instruction, and participants in experimental conditions received an intervention alerting them to an ingroup bias in judgment.

Method

Participants. Participants in Study 1a were University of Virginia (UVA) undergraduates who completed the study for a gift card or course credit. We originally targeted a sample of 652. This sample would provide greater than 80% power at detecting a between-subjects effect of $d = 0.22$, which was the size of the smaller criterion bias found in a pilot study (see supplemental material). Results from this initial sample were inconclusive, so we collected as much additional data as possible during the subsequent semester. To account for inflated Type I errors following multiple rounds of data analysis, we report $p_{\text{augmented}}$ (Sagarin, Ambler, & Lee, 2014). The final sample had 929 participants ($M_{\text{Age}} = 18.9$, $SD = 1.2$, 62.5% female, 59.4% White). See https://osf.io/bm5yk/ for preregistration of materials, https://osf.io/r4xvk/ for analysis plan, and https://osf.io/dpfyv/ for final data collection strategy.

Participants in Study 1b were recruited through an online survey company. We planned for 1,200 participants, which would provide greater than 93% power at detecting at least suggestive evidence ($p < .05$; Benjamin et al., 2018) for a between-subjects effect of $d = 0.20$. In the final sample, 1,223 participants provided data and passed the attention check ($M_{\text{Age}} = 42.4$, $SD = 13.0$, 72.6% female, 63.3% White), and 1,186 of those reported being Democrat or

Republican. See https://osf.io/2dpmx/ for the study's preregistration.

We report all measures, manipulations, and exclusion criteria. Materials, data, analysis syntax, and the supplemental material are available at https://osf.io/mqdga/.

Procedure. Participants in Study 1b completed four study components in the following order: Participants first received the bias-reduction intervention, then completed the academic JBT, followed by items measuring perceptions of JBT performance and explicit attitudes, and finally a demographics survey. Participants in Study 1a completed those same components and a survey about differences among applicants, which followed the JBT, and measures of implicit attitudes, which followed the explicit attitude items.

Experimental conditions. Before completing the JBT, participants were randomly assigned to the Control or Bias Warning condition. Participants in the Control condition received no additional instructions. In Study 1a, participants in the Bias Warning condition were alerted to a bias favoring applicants from one's own university and were asked to avoid showing that bias. In Study 1b, participants in the Bias Warning condition received the same manipulation but about favoring applicants from one's own political party. In Study 1b, participants read the following:

> In addition to differing on their qualifications, candidates will differ in political affiliation. Decision makers are frequently too easy on some applicants and too tough on others. Prior research suggests that decision makers are easier on candidates from their own political party and tougher on candidates from other political parties. Can you be fair toward all applicants and not be biased by applicants' political party? When you make your accept and reject decisions, be as fair as possible. Please tell yourself quietly that you will be fair and avoid favoring applicants from your own political party over applicants from another political party. When you are done, please type this strategy in the box below.

Participants then wrote that they would be fair in a text box. The Bias Warning manipulation was the same for Study 1a, except "university" replaced "political party."

Prior evidence found that this manipulation reduced bias in criterion on a JBT (Axt & Nosek, 2018). For simplicity, we refer to the manipulation as a "bias warning," but the intervention is multifaceted in that it (a) names a specific social dimension known to bias judgment, (b) identifies the direction of the bias, and (c) actively asks participants to avoid showing the bias. We included each of these components in an attempt to maximize the intervention's effectiveness, although we found evidence in other research that removing the text asking participants to avoid bias does not reduce the effect of the manipulation (Axt, 2017).

Academic decision-making task. Participants made accept or reject decisions for an academic honor society. Each applicant had four pieces of academic information: Science GPA (scale of 1-4), Humanities GPA (1-4), letter of recommendation quality (poor, fair, good, excellent), and interview score (1-100). Participants were instructed to accept approximately half of the applicants.

Half of the applications were relatively more qualified and half were less qualified. To determine qualification, each piece of academic information was converted to a 1 to 4 scale. The GPAs already had a maximum score of 4. Recommendation letters were scored *poor* = 1, *fair* = 2, *good* = 3, and *excellent* = 4, and interview scores were divided by 25. The four scores were summed to determine each applicant's level of qualification. Less qualified applicants summed to 13 and more qualified applicants to 14.

In all studies, applicants were paired with equal numbers of male and female faces depicting White, smiling targets. These faces were pretested and divided into two groups differing in physical attractiveness ($d = 2.64$; Axt et al., 2018). In Study 1a, applicants were also depicted with a logo from UVA or a rival school, the University of North Carolina (UNC). Instructions stated that UVA and UNC are equally rigorous, so academic qualifications from both schools should be weighed equally.

In all other studies, applicants varied in both physical attractiveness and political identity. Political identity was depicted by a logo of the Democratic or Republican parties. Participants were reminded of the affiliations for each logo.

In all studies, the JBT contained 64 unique applications, with eight trials (four males, four females) for each combination of qualification, attractiveness, and ingroup membership (eight more physically attractive and more qualified Democrats, eight less physically attractive and more qualified Democrats, etc.). Before evaluating applicants, participants completed an encoding phase where each applicant was shown for 1 s in a random order, although this was removed from Study 1b to save time. Evaluations were made with no time limit.

Participants in Study 1a were assigned to one of two JBT orders. In each, the faces paired with either UVA or UNC were predetermined, but the face-school pairings were randomly assigned to applications during each study session. Across orders, each application was equally likely to be assigned to either a more or less physically attractive face and to be depicted as from UVA versus UNC. The online participants in all other studies were assigned to one of 12 study orders, with each application being equally likely to be assigned to a more versus less physically attractive face or depicted as a Democrat or Republican.

Perceived differences among applicants. Following the JBT, participants in Study 1a rated how different (1 = "not different at all," 5 = "extremely different") applicants were on five dimensions: university affiliation, gender, race, physical attractiveness, and facial expression. These items were not in our analysis plan but were added for exploratory purposes.

Perceptions of performance, explicit attitudes and implicit attitudes. Participants completed four items measuring perceived and desired performance for each social category in the JBT. Each item used a 7-point scale (e.g., –3 = "I was extremely easier on UNC applicants and extremely tougher on UVA applicants; +3 = "I was extremely easier on UVA applicants and extremely tougher on UNC applicants"), with a neutral response indicating equal treatment.

Participants also completed two 7-point explicit preference measures, one for each social category (e.g., –3 = "I strongly prefer less physically attractive people to more physically attractive people, +3 = "I strongly prefer more physically attractive people to less physically attractive people"), with a neutral response indicating no preference.

Participants in Study 1a completed two 7-block evaluative Implicit Association Tests (IAT; Greenwald, McGhee, & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2007) measuring evaluations toward more versus less physically attractive people and identification with UVA versus UNC. See the supplemental material for information procedures for implicit association measures. IATs were completed in a random order and scored with the *D* algorithm (Greenwald, Nosek, & Banaji, 2003).

Demographics. Participants in Study 1a completed a five-item demographics survey. Participants in Study 1b completed a demographics survey reporting political identification, age, gender, and race. For political identification, participants first reported their political party (Democrat, Republican, Independent, Libertarian, Green, Other, Do not know). If participants selected something other than Democrat or Republican, they completed an item asking, if they had to choose, whether they identified more with Democrats or Republicans (participants also had the option to not answer). In Studies 1b to 4 and Study S1, we defined Democrats and Republicans as either those selecting that party identification immediately or those selecting that party in the forced choice. Other evidence suggests these groups are similar in political judgment (Hawkins & Nosek, 2012), and combining them maximized statistical power. Data were analyzed by whether applicants were from the same or opposing political party.

Results

In all studies, participants were excluded from analysis for accepting less than 20% or more than 80% of applicants or for accepting or rejecting every applicant from any social group (Axt, Ebersole, & Nosek, 2016; Axt et al., 2018). Fifteen participants (1.6%) were excluded based on these criteria in Study 1a and 286 participants (24.1%) in Study 1b.[1]

In both studies, accuracy on the JBT (accepting more qualified and rejecting less qualified applicants) was above chance (Study 1a: $M = 70.1\%$, $SD = 7.0$; Study 1b: $M = 62.2\%$, $SD = 9.7$) and the average acceptance rate was close to 50% (Study 1a: $M = 52.9\%$, $SD = 10.1$; Study 1b: $M = 53.4\%$, $SD = 13.6$).

Criterion bias in decision making. For Study 1a, the primary analysis was a 2 (Attractiveness: More vs. Less physically attractive) by 2 (School: UVA vs. UNC) by 2 (Condition: Bias Warning vs. Control) mixed-measures analysis of variance (ANOVA) on criterion for applicants from each combination of school and physical attractiveness. This analysis revealed main effects of physical attractiveness, $F(1, 912) = 130.58$, $p < .001$, $\eta_p^2 = .125$, 95% confidence interval [CI] = [.09, .17], and school, $F(1, 912) = 8.71$, $p = .003$, $\eta_p^2 = .009$, 95% CI = [.001, .03], $p_{augmented} = [.05, .0502]$, with lower criterion for more versus less physically attractive applicants and for applicants from UVA versus UNC. There was no evidence of a main effect of condition, $F(1, 912) = 0.42$, $p = .516$, $\eta_p^2 < .001$, 95% CI = [0, .01]. See Figure 1 for criterion values in each condition for each combination of school affiliation and attractiveness, and see Table 1 for means and standard deviations in each condition for all studies.

A school by condition interaction could indicate that being warned of a school-affiliation bias reduced the bias favoring applicants from one's own university. The school by condition interaction was consistent with this prediction but did not provide strong evidence, $F(1, 912) = 3.03$, $p = .082$, $\eta_p^2 = .003$, 95% CI = [0, .02], $p_{augmented} = [.082, .119]$. The main effect of school was larger in the Control, $F(1, 424) = 9.25$, $p = .002$, $\eta_p^2 = .021$, 95% CI = [.003, .06], $p_{augmented} = [.05, .0501]$, than in the Bias Warning, $F(1, 488) = .87$, $p = .351$, $\eta_p^2 = .002$, condition.

A reliable attractiveness by condition interaction could indicate that being warned of a school-affiliation bias reduced the criterion bias favoring more physically attractive applicants, which would suggest that being warned of a specific bias reduced other biases. The attractiveness by condition interaction was small but suggestive, $F(1, 912) = 5.13$, $p = .024$, $\eta_p^2 = .006$, 95% CI = [0, .02], $p_{augmented} = [.051, .058]$. The main effect of attractiveness was larger in the Control, $F(1, 424) = 80.37$, $p < .001$, $\eta_p^2 = .159$, 95% CI = [.10, .22], than in the Bias Warning, $F(1, 488) = 48.96$, $p < .001$, $\eta_p^2 = .091$, 95% CI= [.05, .14], condition.

Unrelated to key hypotheses, neither the school by attractiveness interaction, $F(1, 912) = 3.72$, $p = .054$, $\eta_p^2 = .004$, 95% CI = [0, .016], nor the school by attractiveness by condition interaction, $F(1, 912) = 2.23$, $p = .136$, $\eta_p^2 = .002$, 95% CI = [0, .013], produced strong effects.

Study 1a results were consistent with the hypothesis that being warned of one bias could reduce a second bias, but evidence was generally weak.

In Study 1b, the primary analysis was a 2 (Attractiveness: More vs. Less physically attractive) by 2 (Political Party: Ingroup vs. Outgroup) by 2 (Condition: Bias Warning vs. Control) mixed-measures ANOVA on criterion for applicants from each combination of political party and physical attractiveness. This analysis revealed main effects of physical

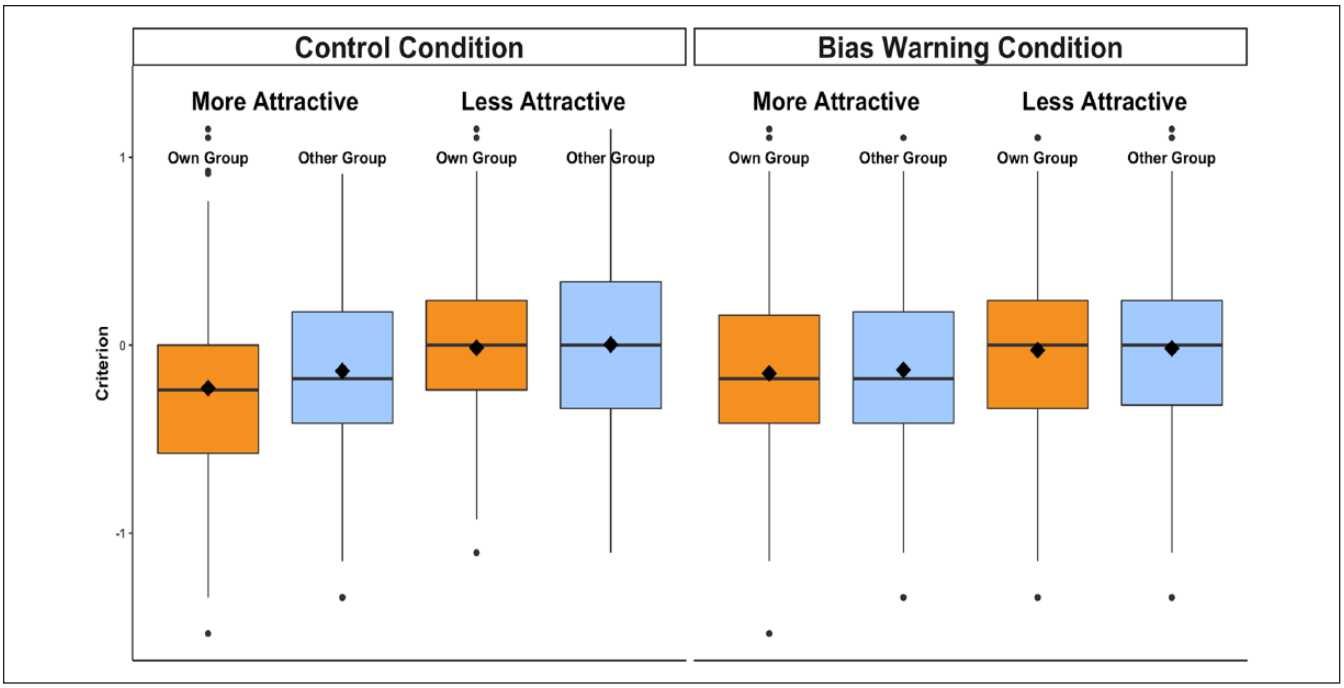

Figure 1. Box plots of criterion values in Study 1a for each experimental condition.
Note. Diamond (♦) denotes mean. Own group were applicants from the decision maker's university, and other group were applicants from another university. In the bias warning condition, participants were asked to avoid favoring applicants from their own university and encouraged to be fair toward own and other group applicants.

attractiveness, $F(1, 898) = 105.77, p < .001$, $\eta_p^2 = .105$, 95% CI= [.07, .14], and political ingroup, $F(1, 898) = 65.82$ $p < .001$, $\eta_p^2 = .068$, 95% CI = [.04, .10], indicating lower criterion for more physically attractive applicants and applicants from one's own political party. There was no evidence of a main effect of condition, $F(1, 898) = 0.59$, $p = .441$, $\eta_p^2 = .001$, 95% CI = [0, .008]. See Figure 2 for criterion values in each condition.

In Study 1b, the political group by condition interaction was reliable, $F(1, 898) = 12.19$, $p = .001$, $\eta_p^2 = .013$, 95% CI = [.003, .03]. The main effect of school was larger in the Control, $F(1, 466) = 55.78, p < .001$, $\eta_p^2 = .107$, 95% CI = [.06, .16], than in the Bias Warning, $F(1, 432) = 14.18$, $p < .001$, $\eta_p^2 = .032$, 95% CI = [.01, 07], condition. Warning participants of a potential political bias reduced the expression of that bias.

However, there was no evidence of an attractiveness by condition interaction, $F(1, 898) = 0.93$, $p = .336$, $\eta_p^2 = .001$, 95% CI = [0, .009]. The main effect of attractiveness in the Control condition, $F(1, 466) = 44.65, p < .001$, $\eta_p^2 = .087$, 95% CI = [.04, .14], was, if anything, slightly smaller than in the Bias Warning condition, $F(1, 432) = 61.70, p < .001$, $\eta_p^2 = .125$, 95% CI = [.07, .18]. Warning participants of a potential political ingroup bias did not reduce expression of the attractiveness bias.

There was no strong evidence for either a political group by attractiveness interaction, $F(1, 898) = 0.10$, $p = .755$, $\eta_p^2 < .001$, or a political group by attractiveness by condition interaction, $F(1, 898) = .97$, $p = .325$, $\eta_p^2 = .001$. Study 1b suggested that being warned of a potential bias for one social category reduced bias for that category but not another social category.

Attitudes, perceived performance, and desired performance. Given similarity with past work (Axt et al., 2018), we review attitude and performance analyses in the aggregate and provide details for individual studies in the supplemental material. Means and standard deviations for implicit and explicit attitudes for all studies are presented in Table 2. The one exception is for implicit attractiveness attitudes, which were only completed in Study 1a (Control IAT $M = 0.74$, $SD = 0.34$; Awareness IAT $M = 0.74$, $SD = 0.33$). As in previous work (Axt et al., 2018), participants showed robust explicit and implicit preferences for more over less physically attractive people (implicit $d = 2.23$ in Study 1a; explicit aggregate $d = 0.72$), explicit preferences for members of their ingroup (aggregate $d = 1.17$), and greater implicit associations between the self and their political party (aggregate $d = 0.88$). Across studies, we found no effects that warning participants of a specific bias changed implicit associations compared with the control condition (Hedges's $g = –.02$, 95% CI = [–.08, .05]), but we did observe small effects on explicit preferences (Hedges's $g = .05$, 95% CI = [.02, .09]). Participants expressed slightly weaker ingroup and attractiveness preferences when they were warned of a potential bias in judgment for that social dimension.

Table 1. Sample Sizes, Criterion Means, and Standard Deviation for All Studies.

	More attractive		Less attractive	
	Own group, *M* (*SD*)	Other group, *M* (*SD*)	Own group, *M* (*SD*)	Other group, *M* (*SD*)
Study 1a condition				
Control (*N* = 425)	−0.23 (0.43)	−0.14 (0.45)	−0.01 (0.43)	0.003 (0.45)
Bias Warning (*N* = 489)	−0.15 (0.42)	−0.13 (0.42)	−0.03 (0.44)	−0.02 (0.43)
Study 1b condition				
Control (*N* = 467)	−0.30 (0.59)	−0.02 (0.62)	−0.16 (0.62)	0.11 (0.66)
Bias Warning (*N* = 433)	−0.25 (0.57)	−0.15 (0.60)	−0.09 (0.58)	0.04 (0.61)
Study 2a condition				
Control (*N* = 462)	−0.20 (0.49)	−0.04 (0.50)	−0.05 (0.53)	0.05 (0.51)
Politics Bias Warning (*N* = 461)	−0.12 (0.47)	−0.08 (0.45)	−0.002 (0.46)	0.03 (0.45)
Attractiveness Bias Warning (*N* = 469)	−0.10 (0.51)	−0.04 (0.48)	−0.10 (0.48)	−0.04 (0.49)
Dual Bias Warning (*N* = 474)	−0.07 (0.47)	0.01 (0.47)	−0.04 (0.49)	0.01 (0.49)
Study 2b condition				
Control (*N* = 362)	−0.17 (0.52)	−0.03 (0.51)	−0.03 (0.50)	0.07 (0.53)
Politics Bias Warning (*N* = 383)	−0.17 (0.48)	−0.12 (0.48)	−0.05 (0.50)	0.02 (0.46)
Attractiveness Bias Warning (*N* = 379)	−0.11 (.47)	−0.01 (.51)	−0.04 (0.50)	0.05 (0.51)
Dual Bias Warning (*N* = 395)	−0.12 (0.49)	−0.06 (0.49)	−0.07 (0.50)	0.01 (0.49)
Study 3 condition				
Control (*N* = 586)	−0.17 (0.51)	−0.03 (0.51)	−0.04 (0.49)	0.12 (0.50)
Dual Bias Warning (*N* = 604)	−0.09 (0.50	−0.01 (0.49)	−0.06 (0.48)	0.02 (0.49)
EEOC Bias Warning (*N* = 577)	−0.14 (0.48)	−0.02 (0.48)	−0.05 (0.48)	0.05 (0.52)
Study 4 condition				
Control (*N* = 546)	−0.16 (0.52)	−0.04 (0.51)	−0.02 (0.52)	0.06 (0.51)
General Bias Warning (*N* = 618)	−0.17 (0.50)	−0.06 (0.50)	−0.01 (0.50)	0.08 (0.50)
Study S1 condition				
Control (*N* = 416)	−0.16 (0.49)	−0.03 (0.49)	−0.04 (0.47)	0.09 (0.52)
Politics Prejudice Warning (*N* = 334)	−0.12 (0.47)	−0.08 (0.51)	0.02 (0.47)	0.001 (0.49)
Attractiveness Prejudice Warning (*N* = 432)	−0.09 (0.51)	0.005 (0.49)	−0.07 (0.47)	−0.01 (0.48)
Dual Prejudice Warning (*N* = 358)	−0.06 (0.47)	−0.02 (0.45)	−0.01 (0.49)	−0.005 (0.48)

Note. EEOC = Equal Employment Opportunity Commission.

Means and standard deviations for perceived and desired performance are presented in Table 2; Table 3 presents the percentage of participants who reported having shown no bias on the task or wanting to show no bias for both physical attractiveness and ingroup status. Across studies, participants perceived having favored more physically attractive applicants (aggregate $d = 0.18$) and members of their own political or university group (aggregate $d = 0.35$). On average, participants reported a moderate desire to favor ingroup members (aggregate $d = 0.32$) and a very small desire to favor more physically attractive applicants (aggregate $d = 0.04$).

Nevertheless, a large majority of participants reported a desire to be unbiased on the JBT (attractiveness: 91.8%; ingroup 86.3%) and a perception of having been unbiased (attractiveness: 79.3%; ingroup 82.0%). This suggests that most participants viewed these biases as socially unacceptable in the context of an academic evaluation, but a few felt free to express ingroup favoritism. In line with actual JBT behavior, bias warnings were associated with small changes in desired performance (Hedges's $g = .10$, 95% CI = [.05, .15]) and perceived performance (Hedges's $g = .18$, 95% CI = [.14, .23]) such that perceived and desired performance indicated more equal treatment after being warned about the potential for a specific bias.

Finally, across studies, correlations with JBT performance replicated prior work (Axt et al., 2016; Axt et al., 2018): Criterion bias was weakly but positively associated with explicit attitudes ($r = .16$, 95% CI = [.12, .21]), perceived performance ($r = .29$, 95% CI = [.22, .36]), desired performance ($r = .22$, 95% CI = [.15, .29]), and implicit associations ($r = .12$, 95% CI = [.10, .14]).

Discussion

In control conditions, participants displayed two simultaneous biases: favoritism toward more physically attractive people and toward ingroup members. Warning participants of a potential university or political bias, and asking them to avoid displaying such biases, reduced that bias, though very weakly in Study 1a. Warning participants of a potential

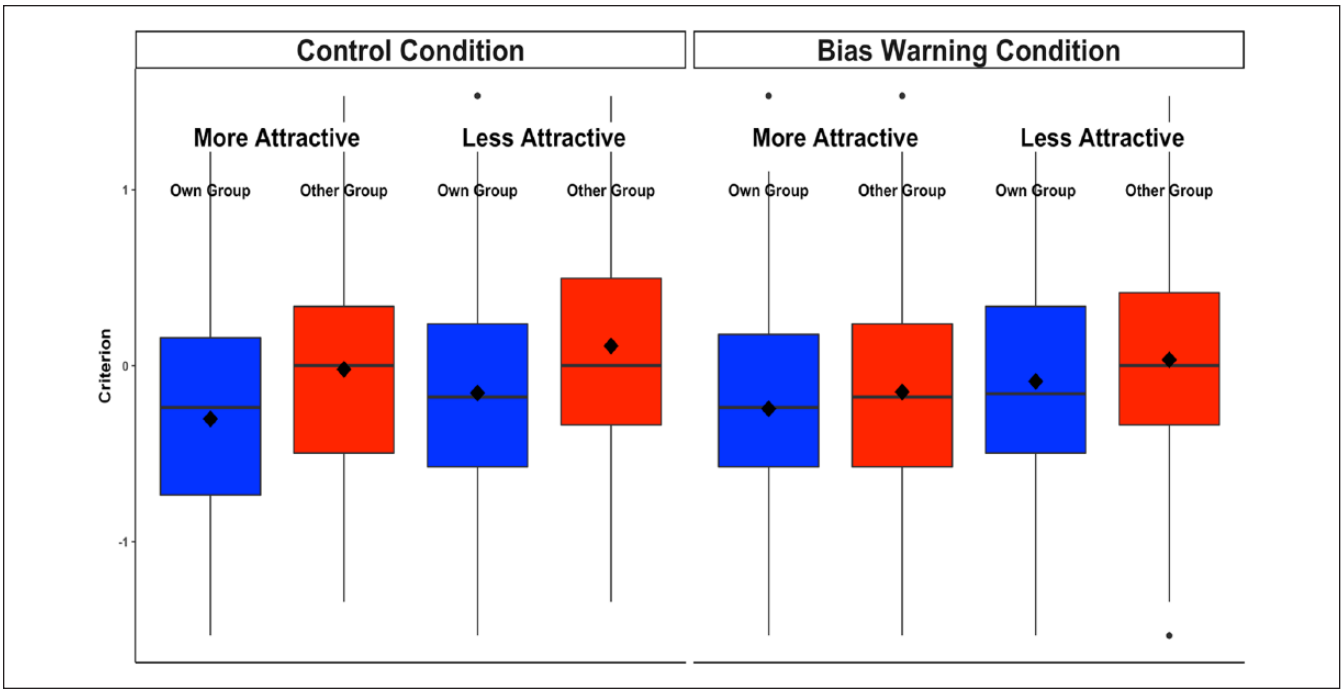

Figure 2. Box plots of criterion values in Study 1b for each experimental condition.
Note. Diamond (♦) denotes mean. Own group were applicants from the decision maker's political party, and other group were applicants from the other political party. In the bias warning condition, participants were asked to avoid favoring applicants from their own political party and encouraged to be fair toward own and other group applicants.

university or political bias reduced attractiveness bias weakly in Study 1a and not at all in Study 1b.

Despite relatively high-powered tests, these results do not definitively support or reject the possibility that an intervention warning about bias in one category influences a second, unmentioned social category bias in judgment. We sought more evidence by replicating and extending these findings. In Studies 2a and 2b, we replicated Study 1b and extended it into a 2 × 2 design, making warning participants of none, one, or both of the social judgment biases.

Studies 2a and 2b

Method

Participants. Participants in Studies 2a to 4 came from the Project Implicit research pool. In Study 2a, we selected Americans who reported being at least slightly liberal or conservative and targeted a sample of 400 per condition. This sample would provide over 90% power at detecting a between-subjects effect of $d = 0.24$, which was the size of the impact of the warning for politics bias in Study 1b. Exclusion rates were overestimated, leading to a final sample larger than anticipated. In total, 2,106 participants (M_{Age} = 37.5, SD = 14.9, 61.9% female, 72.8% White) provided data, and 1,993 reported being a Democrat or Republican. See https://osf.io/c6vjr/ for the study's preregistration.

In Study 2b, we selected Americans who reported being at least slightly liberal or conservative and targeted a sample of 357 for each condition. This sample provided over 80% power at detecting a between-subjects effect of $d = 0.21$, which was the average effect for the impact of warning about a specific bias for reducing that bias on the JBT in Studies 1b to 2a. In total, 1,744 participants (M_{Age} = 32.8, SD = 14.6, 69.5% female, 73.4% White) provided data, and 1,623 reported being a Democrat or Republican. See https://osf.io/ynhup/ for the study's preregistration.

Procedure. In both studies, participants completed five components in the following order: bias warning intervention, academic JBT, self-report items of JBT performance, explicit attitudes and political identity, and a measure of implicit political identification.

In Study 2b, immediately after the JBT, participants also completed five items assessing perceptions of differences among applicants' attractiveness, gender, race, political party, and GPA. Our preregistered analysis plan noted that we would only analyze these items if we replicated a three-way interaction found in Study 2a, which we did not. As a result, we did not analyze these items and do not discuss them further.

Experimental conditions. In both studies, participants were randomly assigned to one of four conditions: Control, Politics Bias Warning, Attractiveness Bias Warning, or Dual Bias Warning. In the Control condition, participants received no additional instructions. In the Politics Bias Warning condition, participants read the same manipulation as Study 1b. In the Attractiveness Bias Warning condition, participants read

Table 2. Means (and Standard Deviations) for Attitude and Performance Measures.

	Attractiveness			Ingroup		
	Exp. preference	Perc. performance	Des. performance	Exp. preference	Imp. association	Perc. performance	Des. performance
Study 1a							
Control	1.01 (0.82)	0.39 (0.69)	0.07 (0.47)	0.72 (0.91)	0.41 (0.32)	0.19 (0.49)	0.17 (0.54)
Bias Warning	1.07 (0.85)	0.31 (0.60)	0.08 (.43)	0.66 (0.88)	0.39 (0.29)	0.10 (0.40)	0.06 (0.36)
Study 1b							
Control	0.44 (1.03)	0.05 (0.83)	0.04 (0.76)	1.22 (1.28)		0.32 (0.87)	0.32 (0.81)
Bias Warning	0.38 (0.95)	0.09 (0.77)	0.09 (0.61)	1.03 (1.16)		0.20 (0.73)	0.24 (0.74)
Study 2a							
Control	0.71 (0.98)	0.20 (0.66)	0.05 (0.43)	1.38 (1.17)	0.35 (0.44)	0.31 (0.75)	0.22 (0.65)
Politics Bias Warning	0.62 (0.96)	0.18 (0.67)	0.04 (0.44)	1.41 (1.21)	0.42 (0.47)	0.10 (0.48)	0.12 (0.49)
Attractiveness Bias Warning	0.62 (0.97)	−0.02 (0.61)	−0.02 (0.48)	1.37 (1.15)	0.35 (0.46)	0.18 (0.58)	0.16 (0.56)
Dual Bias Warning	0.51 (0.90)	0.002 (0.57)	−0.03 (0.37)	1.19 (1.21)	0.38 (0.46)	0.14 (0.52)	0.06 (0.40)
Study 2b							
Control	0.70 (1.01)	0.15 (0.75)	0.02 (0.49)	1.34 (1.17)	0.38 (0.44)	0.29 (0.68)	0.22 (0.65)
Politics Bias Warning	0.65 (1.00)	0.10 (0.63)	0.02 (0.32)	1.29 (1.21)	0.34 (0.49)	0.19 (0.64)	0.15 (0.58)
Attractiveness Bias Warning	0.60 (0.95)	0.05 (0.62)	0.02 (0.48)	1.29 (1.15)	0.36 (.46)	0.27 (0.68)	0.20 (0.63)
Dual Bias Warning	0.67 (0.90)	0.07 (0.59)	0.04 (0.48)	1.27 (1.14)	0.35 (0.44)	0.17 (0.56)	0.19 (0.56)
Study 3							
Control	0.65 (0.89)	0.15 (0.67)	0.002 (0.39)	1.67 (1.14)	0.38 (0.46)	0.36 (0.77)	0.29 (0.73)
Dual Bias Warning	0.67 (0.94)	0.07 (0.59)	−0.003 (0.39)	1.67 (1.15)	0.41 (0.43)	0.19 (0.63)	0.20 (0.62)
EEOC Bias Warning	0.67 (0.92)	0.11 (0.61)	−0.03 (0.36)	1.61 (1.13)	0.42 (0.46)	0.27 (0.66)	0.24 (0.66)
Study 4							
Control	0.77 (0.97)	0.12 (0.72)	0.004 (0.41)	1.60 (1.11)	0.40 (0.46)	0.29 (0.72)	0.22 (0.67)
General Bias Warning	0.75 (0.93)	0.14 (0.70)	0.03 (0.40)	1.54 (1.13)	0.38 (0.45)	0.25 (0.64)	0.23 (0.65)
Study S1							
Control	0.62 (0.84)	0.05 (0.55)	−0.01 (0.39)	1.44 (1.19)	0.38 (0.45)	0.28 (0.69)	0.21 (0.62)
Politics Prejudice Warning	0.62 (0.94)	0.14 (0.58)	0.04 (0.51)	1.48 (1.15)	0.37 (0.46)	0.17 (0.55)	0.14 (0.52)
Attractiveness Prejudice Warning	0.56 (0.83)	0.03 (0.48)	−0.02 (0.39)	1.43 (1.20)	0.37 (0.46)	0.20 (0.55)	0.23 (0.63)
Dual Prejudice Warning	0.54 (0.88)	−0.02 (0.39)	−0.01 (0.39)	1.34 (1.15)	0.40 (0.47)	0.15 (0.57)	0.16 (0.56)

Note. For ingroup items, higher values mean more preference, stronger implicit identification, or more perceived/desired favoritism for members of one's own group (university in Study 1a, political party in all other studies). Exp. Preference = explicit preference item; Perc. Performance = perceived performance item; Des. Performance = desired performance item; Imp. Association = implicit association measures (IAT in Study 1a, BIAT in all other studies); IAT = Implicit Association Tests; BIAT = Brief Implicit Association Test; EEOC = Equal Employment Opportunity Commission.

an updated version of that manipulation replacing any mention of political ingroup bias with a bias favoring more physically attractive people. In the Dual Bias Warning condition, participants read the relevant text from both the Attractiveness and Politics Bias Warning conditions. See supplemental material for full text.

Academic decision-making task. Participants completed the same JBT as Study 1b.

Perceptions of performance, explicit attitudes, and political identity. Participants completed the same explicit attitude and performance measures, as well as the same measure of self-reported political identification as Study 1b.

Implicit identification. Participants completed a four-block, self-focal Brief Implicit Association Test (Sriram & Greenwald, 2009) measuring identification with Democrats versus Republicans.

Results

In all, 127 (6.4%) participants in Study 2a and 104 (6.4%) participants in Study 2b were excluded from analysis. In both studies, overall JBT accuracy was above chance (Study 2a: $M = 68.4\%$, $SD = 8.4$; Study 2b: $M = 66.7\%$, $SD = 9.0$) and average acceptance rate was close to 50% (Study 1a: $M = 51.2\%$, $SD = 12.0$; Study 1b: $M = 51.8\%$, $SD = 12.1$).

Criterion bias in decision making. For Studies 2a and 2b, primary analyses focused on a 2 (Applicant Attractiveness: More vs. Less physically attractive) by 2 (Applicant Politics: Own party vs. Other party) by 2 (Attractiveness Condition: Bias Warning vs. None) by 2 (Politics Condition: Bias

Table 3. Percentage of Participants Reporting a Desire or Perception of Having Behaved Fairly.

	Attractiveness		Ingroup	
	Perceived performance (%)	Desired performance (%)	Perceived performance (%)	Desired performance (%)
Study 1a				
Control	63.3	88.5	81.9	88.0
Bias Warning	69.5	92.0	86.1	92.6
Study 1b				
Control	78.0	84.0	75.9	78.5
Bias Warning	83.1	87.3	86.7	84.1
Study 2a				
Control	75.4	91.1	79.6	83.9
Politics Bias Warning	79.4	92.2	87.1	91.9
Attractiveness bias Warning	82.5	91.9	85.1	87.9
Dual Bias Warning	83.9	93.5	88.5	93.9
Study 2b				
Control	78.3	90.3	77.9	83.7
Politics Bias Warning	78.3	95.0	82.8	89.9
Attractiveness Bias Warning	82.0	92.0	81.5	87.5
Dual Bias Warning	79.9	90.6	87.4	86.3
Study 3				
Control	76.2	94.9	74.5	80.7
Dual Bias Warning	84.5	93.7	81.9	85.6
EEOC Bias Warning	80.8	93.2	78.0	83.8
Study 4 condition				
Control	77.1	92.0	77.8	85.0
General Bias Warning	77.9	91.9	80.4	84.1
Study S1				
Control	84.5	93.3	79.8	84.5
Politics Prejudice Warning	83.1	90.5	84.6	88.6
Attractiveness Prejudice Warning	85.4	92.7	84.1	85.5
Dual Prejudice Warning	85.0	93.4	86.5	89.7

Note. EEOC = Equal Employment Opportunity Commission.

Warning vs. None) mixed-measures ANOVA on criterion for each combination of applicant attractiveness and political identity.

Both studies showed main effects of physical attractiveness such that more physically attractive applicants received lower criterion than less physically attractive applicants—Study 2a: $F(1, 1862) = 47.58, p < .001$, $\eta_p^2 = .025$, 95% CI = [.013, .041]; Study 2b: $F(1, 1515) = 92.90, p < .001$, $\eta_p^2 = .058$, 95% CI = [.037, .082]. There were also main effects of applicant political party such that applicants from one's own party received lower criterion than from the other party—Study 2a: $F(1, 1862) = 57.61, p < .001$, $\eta_p^2 = .030$, 95% CI = [.017, .047]; Study 2b: $F(1, 1515) = 56.10, p < .001$, $\eta_p^2 = .036$, 95% CI = [.020, .056]. Main effects of bias warnings for attractiveness or politics were not reliable ($\eta_p^2 < .002$). See Figure 3 for criterion values in Study 2a and Figure 4 for Study 2b.

If warning about a specific bias reduced that bias, there would be evidence of interactions between applicant attractiveness and attractiveness bias warning condition as well as between applicant political ingroup and politics bias warning condition. Both studies showed an applicant attractiveness by attractiveness bias warning interaction—Study 2a: $F(1, 1862) = 35.60, p < .001$, $\eta_p^2 = .019$, 95% CI = [.009, .033]; Study 2b: $F(1, 1515) = 9.40, p = .002$, $\eta_p^2 = .006$, 95% CI= [.001, .016]; the attractiveness bias was smaller when participants were warned about the potential attractiveness bias (Study 2a: $\eta_p^2 = .0004$; Study 2b: $\eta_p^2 = .027$) than when not (Study 2a: $\eta_p^2 = .086$; Study 2b: $\eta_p^2 = .100$).

There was weak evidence of political ingroup by politics bias warning interactions in Study 2a, $F(1, 1862) = 5.19$, $p = .023$, $\eta_p^2 = .003$, 95% CI = [.00002, .010], and Study 2b, $F(1, 1515) = 3.34, p = .068$, $\eta_p^2 = .002$, 95% CI = [0, .009]. In both cases, the main effect of political ingroup was smaller when participants had been warned about the bias (Study 2a: $\eta_p^2 = .018$; Study 2b: $\eta_p^2 = .024$) than when not (Study 2a: $\eta_p^2 = .043$; Study 2b: $\eta_p^2 = .047$). The evidence for warnings

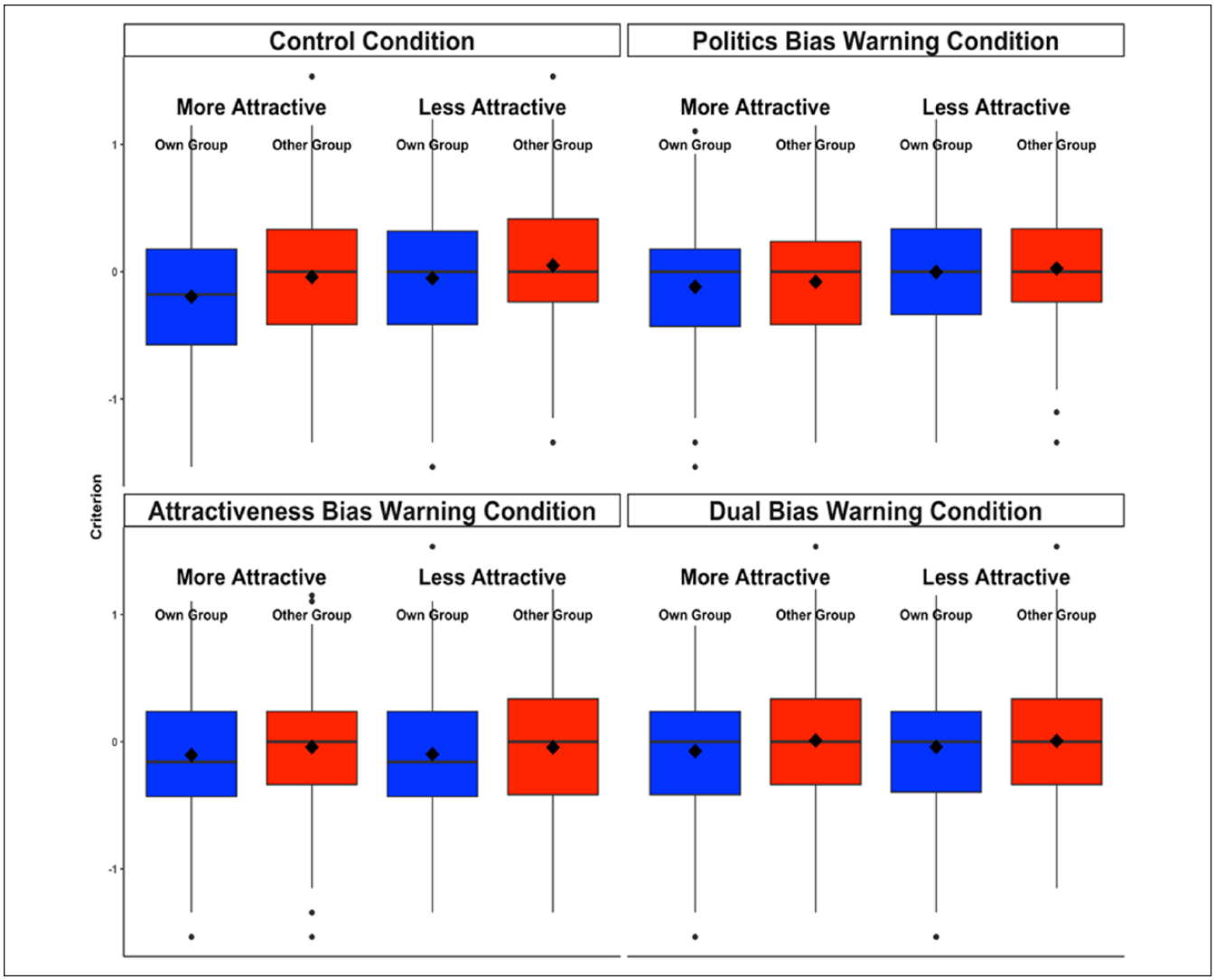

Figure 3. Box plots of criterion values in Study 2a for each experimental condition.
Note. Diamond (♦) denotes mean. Own group were applicants from the decision maker's political party, and other group were applicants from the other political party.

reducing biased judgment was consistent for attractiveness and politics, but stronger for attractiveness.

A central question was whether warning about bias for one category influenced the bias expressed on the other category. If so, we would observe interactions between applicant attractiveness and politics bias warning conditions and between applicant political ingroup status and attractiveness bias warning conditions. In both studies, there was no evidence of interactions between applicant attractiveness and politics bias warning condition—Study 2a: $F(1, 1862) = 0.03, p = .868, \eta_p^2 < .001$; Study 2b: $F(1, 1515) = 0.03, p = .865, \eta_p^2 < .001$—or between applicant political ingroup by attractiveness bias warning condition—Study 2a: $F(1, 1862) = 0.94, p = .333, \eta_p^2 = .001$; Study 2b: $F(1, 1515) = .11, p = .744, \eta_p^2 < .001$.

It is possible that the effectiveness of the bias warning interventions was moderated by whether participants were also warned about the other potential source of bias. This would be indicated by three-way interactions of both bias warning manipulations and the category manipulation (attractiveness or politics). There was no evidence of an interaction between applicant attractiveness and the two warning conditions—Study 2a: $F(1, 1862) = 0.29, p = .593, \eta_p^2 < .001$; Study 2b: $F(1, 1515) = .13, p = .720, \eta_p^2 < .001$. Also, there was no evidence of an interaction between applicant politics, attractiveness bias warning, and politics bias warning in Study 2b, $F(1, 1515) = 0.46, p = .500, \eta_p^2 < .001$. However, there was suggestive evidence of such an interaction in Study 2a, $F(1, 1862) = 7.37, p = .007, \eta_p^2 = .004$, 95% CI = [.0003, .012]. The impact of the politics bias warning manipulation on reducing political bias was stronger when participants were only warned about the politics bias ($\eta_p^2 = .012$) compared with when they were also warned about an attractiveness bias ($\eta_p^2 < .001$). While intriguing, as

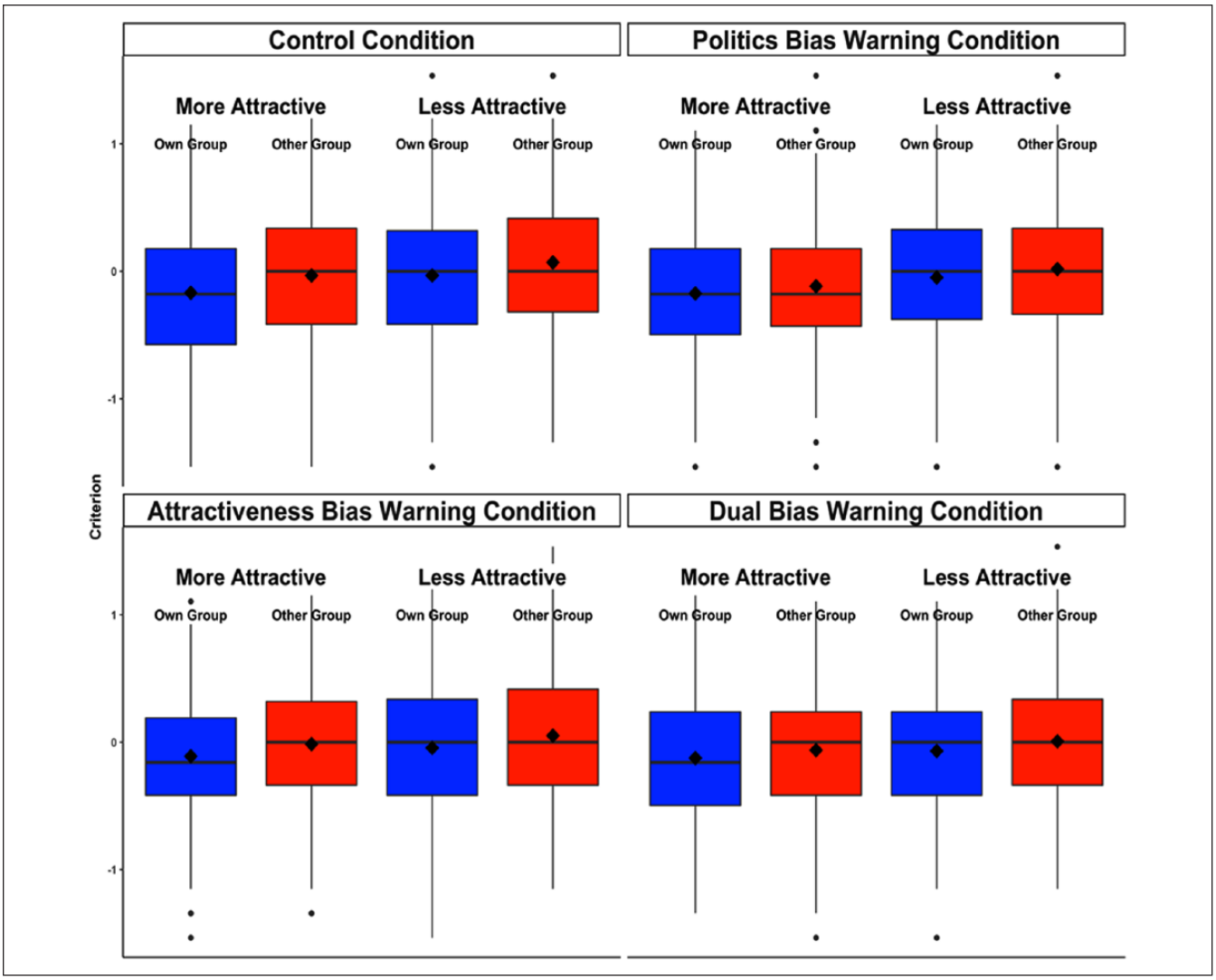

Figure 4. Box plots of criterion values in Study 2b for each experimental condition.
Note. Diamond (♦) denotes mean. Own group were applicants from the decision maker's political party, and other group were applicants from the other political party.

a single instance among multiple tests, this is weak evidence that should be replicated before taken seriously.

No other terms in the ANOVA in either study reached suggestive ($p < .05$) or stronger ($p < .005$) statistical significance (Benjamin et al., 2018). See supplemental material for full reporting.

Discussion

Warning participants about a potential bias in social judgment was effective at reducing the bias, but only if the social category was explicitly identified. Unlike the mixed evidence in Studies 1a and 1b, we observed no evidence in Studies 2a and 2b that warning of a potential bias toward one social category reduced bias toward another social category.

Study 3

The general conclusion of Studies 1a to 2b is that warning about potential biases in judgment decreases biased behavior for the specific social categories mentioned but does not transfer to reduced judgment biases for other, unmentioned social categories. One logical extension of this result is to examine what happens when participants are warned about *many* potential biases. This question has both theoretical relevance in establishing whether there is an upper bound in the number of categories that can be listed in an effective intervention and practical consequences given that existing anti-discriminatory warnings for evaluation (e.g., the EEOC language for protected classes) list many social categories as potentially influencing judgment.

It is possible that raising warning about many potential biases will increase the effectiveness of the intervention, as

perhaps listing many social categories heightens participant vigilance or motivation to be unbiased. Conversely, it is also possible that listing many social categories will decrease the effectiveness of the intervention, as perhaps participants' attention may be less focused on limiting the influence of the relevant social categories responsible for biased judgment in the present context. In Study 3, we tested whether interventions warning participants of multiple biases were as effective as interventions warning participants only about the biases concerning the social categories responsible for bias. Finally, in the "General Discussion" section, we report post hoc analyses of Studies 1a to 2b finding evidence of a gender bias in criterion such that participants had lower criterion for female than male applicants (Hedges's $g = .31$). Study 3 was designed and conducted following these post hoc analyses and as a result included applicant gender in confirmatory analyses.

Method

Participants. In Study 3, we selected Project Implicit participants who reported being American citizens and at least slightly liberal or conservative. We targeted a sample of 581 participants for each experimental condition, which would provide 80% power at detecting the average effect size in Studies 1b to 2a for the impact of warning about a specific bias ($d = 0.165$). In total, 1,963 participants ($M_{Age} = 37.4$, $SD = 15.5$, 70.9% female, 73.7% White) provided data, and 1,873 reported being a Democrat or Republican. See https://osf.io/aktpw/ for the study's preregistration.

Procedure. Participants completed the same measures as in Study 2a with two changes. First, participants also completed items about explicit preferences, desired performance, and perceived performance for male versus female applicants (see supplemental material for descriptive statistics of gender measures).

Second, participants were assigned to one of three experimental conditions. Participants in the Control condition completed the JBT without any additional instructions, and participants in the Dual Bias Warning condition viewed the same intervention as in Studies 2a to 2b. Finally, participants in the EEOC Bias Warning condition read the following:

> In addition to differing on their qualifications, applicants will differ in other ways such as age, attractiveness, citizenship status, color, creed, disability, gender identity and/or expression, genetic information, marital status, national origin, political orientation, race, religion, sex, sexual orientation, status with regard to public assistance, veteran status, or any other characteristic protected by federal, state, or local law. Decision makers are frequently too easy on some applicants and too tough on others. Prior research suggests that decision makers are easier on people of some identities from these categories and harder on people with other identities from these categories.

As in the Dual Bias Warnings condition, participants were asked to be fair toward applicants from all of the listed social categories.

Results

In all, 106 (5.6%) participants were removed from analysis due to our JBT exclusion criteria. JBT accuracy was above chance ($M = 69.0\%$, $SD = 8.6$) and average acceptance rate was close to 50% ($M = 51.1\%$, $SD = 12.2$).

Criterion bias in decision making. Study 3 analyses included gender as a factor in analysis to test whether the EEOC Bias Warning condition would effectively reduce the gender bias found in prior studies. Our primary analysis was then a 2 (Attractiveness: More vs. Less physically attractive) by 2 (Political Party: Ingroup vs. Outgroup) by 2 (Gender: Female vs. Male) by 3 (Experimental condition) ANOVA on criterion values. As expected, there were main effects of gender, $F(1, 1764) = 247.00$, $p < .001$, $\eta_p^2 = .123$, 95% CI = [.10, .15], such that female applicants received a lower criterion than male applicants; attractiveness, $F(1, 1764) = 90.27$, $p < .001$, $\eta_p^2 = .049$, 95% CI= [.03, .07], such that more physically attractive applicants received a lower criterion than less physically attractive applicants; and political party, $F(1, 1764) = 112.99$, $p < .001$, $\eta_p^2 = .060$, 95% CI = [.04, .08]), such that applicants from one's own party received a lower criterion than applicants from the other party.

If any of these main effects interacted with condition, it would suggest that conditions differed in the degree to which these social dimensions affected criterion. There was no reliable interaction between condition and gender, $F(2, 1764) = 0.67$, $p = .512$, $\eta_p^2 = .001$, but there was a reliable interaction between condition and attractiveness, $F(2, 1764) = 11.14$, $p < .001$, $\eta_p^2 = .012$, 95% CI = [.004, .02], and a suggestive but very small interaction between condition and political party, $F(2, 1764) = 3.46$, $p = .032$, $\eta_p^2 = .004$, 95% CI = [0, .01]. None of the three-way interactions involving experimental condition were reliable (all Fs < 1.05, all ps $> .349$; see supplemental material for full ANOVA tables).

We then ran follow-up ANOVAs comparing each of the three condition pairings on the relative impact of attractiveness and political affiliation on criterion, reporting interactions between each social dimension and condition. First, relative to the Control condition, the Dual Bias Warning condition reduced both the impact of physical attractiveness, $F(1, 1188) = 22.13$, $p < .001$, $\eta_p^2 = .018$, 95% CI= [.01, .04], and political party affiliation, $F(1, 1188) = 6.55$, $p = .011$, $\eta_p^2 = .005$, 95% CI = [.001, .02], on criterion. Second, relative to the Control condition, the EEOC Bias Warning condition suggestively reduced the impact of physical attractiveness on criterion, $F(1, 1161) = 6.61$, $p = .010$, $\eta_p^2 = .006$, 95% CI = [.001, .02], but did not affect political affiliation, $F(1, 1161) = 2.14$, $p = .143$, $\eta_p^2 = .002$. Finally, relative to the EEOC Bias Warning condition, there was suggestive

evidence that the Dual Bias Warning condition further reduced the impact of physical attractiveness, $F(1, 1179) = 4.44$, $p = .035$, $\eta_p^2 = .004$, 95% CI = [0, .01], but did not reliably change the impact of political affiliation, $F(1, 1179) = 1.35$, $p = .245$, $\eta_p^2 = .001$.

Analyses within each condition found that no intervention fully removed the impact of attractiveness or political affiliation on criterion, as each condition still showed reliable main effects of attractiveness and political affiliation (all Fs > 5.48, all ps $< .020$; see supplemental material for full reporting and additional preregistered analyses).

Discussion

Replicating Studies 2a and 2b, an intervention that warned participants about biases concerning both physical attractiveness and political affiliation reduced those biases in judgment. A novel intervention that warned about these and other biases (e.g., race, sexual orientation) reduced the attractiveness but not the political affiliation bias. Moreover, the evidence suggests that mentioning many potential sources of bias was less effective at reducing favoritism than only mentioning physical attractiveness and political ingroup affiliation. Furthermore, neither condition reliably affected gender biases in evaluation that favored female applicants, despite gender being identified explicitly in the condition using EEOC language.

Cumulatively, this evidence suggests that interventions warning about biases for many possible categories reduce the effectiveness of the intervention compared with warning specifically about only one or two social categories as the source of potential bias. Just one of the three examined categories showed any evidence of reduced bias in the EEOC condition, and that was weaker than the more focused biased warning condition.

Meta-Analysis of Bias Warnings

To provide more precise estimates of the impact of warning participants to avoid bias on targeted and untargeted biases, we conducted a "mini meta-analysis" (Goh, Hall, & Rosenthal, 2016) of Studies 1a to 2b, Study 3 (excluding the *EEOC* condition), and Study S1 ($N = 1{,}540$, full report in supplemental material). Study S1 was a replication of Study 2a, with the one change being that participants were warned to avoid showing "prejudice" instead of "bias" toward one, both, or neither social group included in the JBT. Study S1 results found similar effects as Studies 2a to 2b, and the study is described further in the "General Discussion" section.

We converted each η_p^2 into a Pearson's correlation (r) such that higher r values mean the bias warning intervention more effectively reduced criterion bias. There was a small but reliable meta-analytic effect that warning about bias concerning a social category was associated with reduced criterion bias toward that category ($r = .095$, 95% CI= [.071, .118], $Z = 7.93$, $p < .001$). However, there was no reliable meta-analytic effect of warning about bias for one social category reducing criterion bias toward the unmentioned social category ($r = .006$, 95% CI = [–.012, .0239], $Z = 0.62$, $p = .533$). See Figure 5 for forest plots. The meta-analytic results provide a relatively unambiguous conclusion—interventions alerting participants to a specific bias for a social category and encouraging fairness in evaluation were effective at reducing social judgment bias toward that category but not toward another, unmentioned social category.

Study 4

Studies 1a to 3 suggest that mentioning the specific social category responsible for bias may be a necessary condition for bias-reduction interventions to be effective, and that the impact of such interventions may be weakened when warning participants about the potential for biases concerning many social categories. This introduces an obvious question—what happens with the same intervention, but instead of highlighting any specific social category or categories, the intervention refers more generally to different "types of applicants"? If mentioning the social category directly is necessary for reducing bias, then this intervention will be ineffective. However, such a "general" intervention could make decision makers attentive to multiple sources of potential bias rather than an intervention directing them to one specific social category (e.g., Sassenberg & Moskowitz, 2005), and potentially at the cost of attending to other social categories. In Study 4, we test this possibility.

Method

Participants. We preselected Americans identifying as liberal or conservative and targeted an average of 581 participants across the two conditions. This sample would provide 80% power at detecting the average effect in Studies 1 to 2b for the impact of warning about a specific bias ($d = 0.165$). In total, 1,300 participants ($M_{Age} = 33.0$, $SD = 15.2$, 69.8% female, 68.0% White) provided data, and 1,249 reported being a Democrat or Republican. See https://osf.io/adbq3/ for the study's preregistration.

Procedure. Participants completed the same measures as in Study 2a. The only change was the contents of the bias-reduction manipulation.

In Study 4, participants were randomly assigned to a Control or General Bias Warning condition. In the Control condition, participants received no additional instructions. In the General Bias Warning condition, participants received the same manipulation as participants in the Politics Bias Warning conditions as Studies 2a and 2b, except any mention of the politics social category was removed. Specifically, participants read,

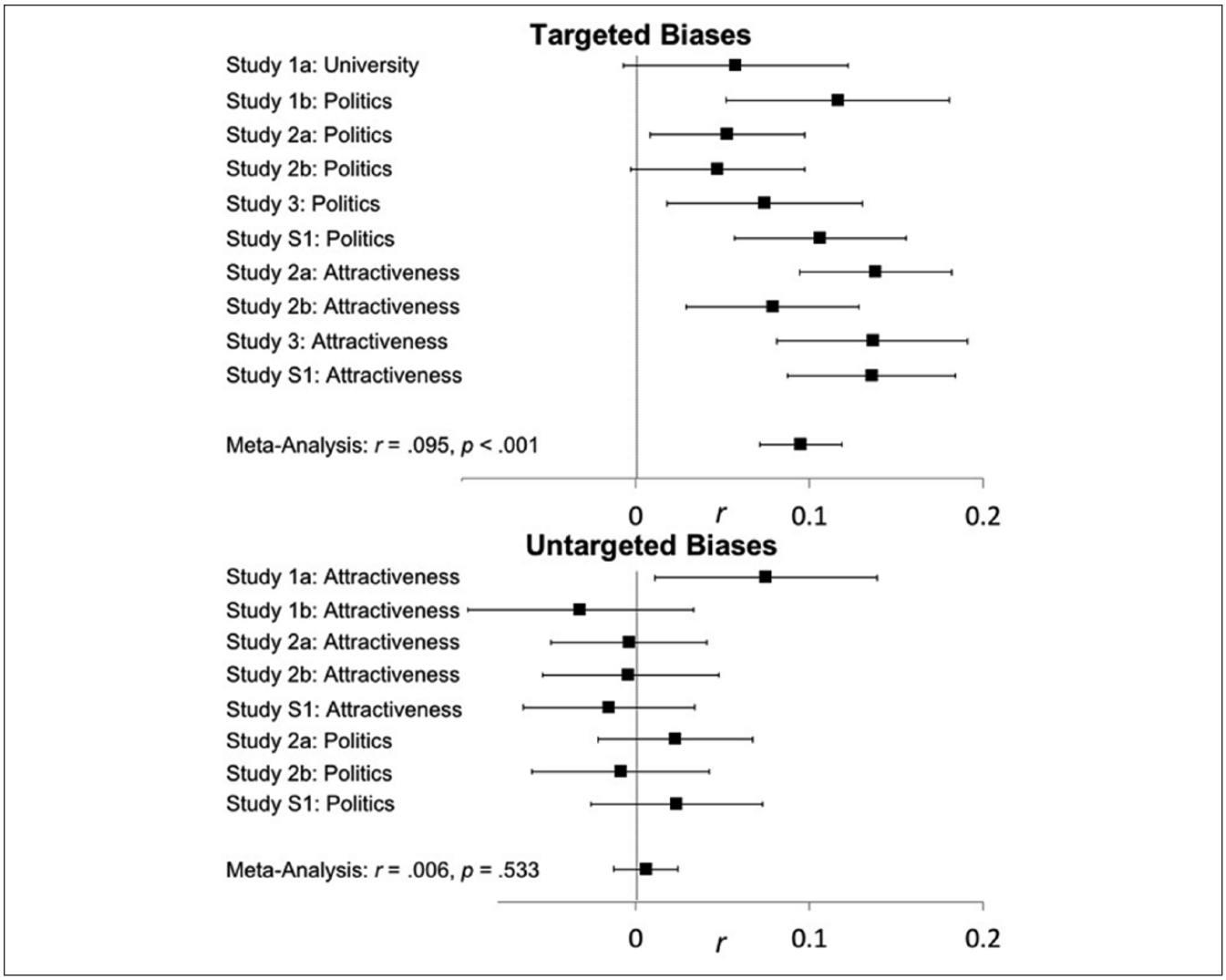

Figure 5. Forest plot and meta-analysis for the impact of bias warnings on targeted and untargeted biases in Studies 1a to 3.
Note. More positive values indicate that the bias warning manipulation reduced criterion bias on that category relative to the control condition. Error bars denote 95% confidence intervals on the effect size.

> In addition to differing on their qualifications, candidates will differ in other ways. Prior research suggests that decision makers are frequently too easy on some types of applicants and too tough on others. Can you be fair toward all applicants? When you make your accept and reject decisions, be as fair as possible. Please tell yourself quietly that you will be fair. When you are done, please type the strategy "I will be fair" in the box below.

Results

In all, 85 (6.8%) participants were excluded from analysis. JBT Accuracy was above chance ($M = 67.6\%$, $SD = 8.5$) and average acceptance rate was close to 50% ($M = 51.3\%$, $SD = 12.5$).

Criterion bias in decision making. The primary analysis was a 2 (Attractiveness: More vs. Less physically attractive) by 2 (Political Party: Ingroup vs. Outgroup) by 2 (Condition: Control vs. General Bias Warning) mixed-measures ANOVA on criterion. This analysis revealed main effects of physical attractiveness, $F(1, 1162) = 159.29$, $p < .001$, $\eta_p^2 = .120$, 95% CI= [.09, .16], and political ingroup, $F(1, 1162) = 56.60$, $p < .001$, $\eta_p^2 = .046$, 95% CI = [.03, .07], with lower criterion for more physically attractive applicants and applicants from one's own political ingroup. There was no evidence of a main effect of condition, $F(1, 1162) < 0.001$, $p = .996$, $\eta_p^2 < .001$.

Of primary interest were the attractiveness by condition and politics by condition interactions, which could indicate that the general bias warning manipulation affected evaluation of more versus less physically attractive applicants or applicants from one's own versus the other political party. There was no evidence for interactions between applicant

attractiveness and condition, $F(1, 1162) = 1.74$, $p = .188$, $\eta_p^2 = .001$, or between applicant political ingroup and condition, $F(1, 1162) = .32$, $p = .573$, $\eta_p^2 = .003$. The main effects of politics were similar in the Control, $F(1, 545) = 26.38$ $p < .001$, $\eta_p^2 = .046$, 95% CI = [.018, .085], and the General Bias Warning, $F(1, 617) = 30.33$, $p < .001$, $\eta_p^2 = .047$, 95% CI = [.020, .083], condition. The main effect of attractiveness in the Control condition, $F(1, 545) = 58.91$, $p < .001$, $\eta_p^2 = .098$, 95% CI = [.055, .146], was, if anything, slightly smaller than in the General Bias Warning condition, $F(1, 617) = 104.26$, $p < .001$, $\eta_p^2 = .145$, 95% CI = [.097, .195].

Participants warned about a general potential for bias did not show reduced bias on either social category, and the impact of general warnings about bias favoring attractiveness ($r = –.04$, 95% CI = [–.10, .02]) or the political ingroup ($r = –.004$, 95% CI = [–.06, .05]) fell outside the 95% CI for the impact of warning about specific biases reported in the meta-analysis of Studies 1a to 3 and Study S1. This evidence is consistent with the interpretation that warning about bias in a specific social category is necessary for reducing social judgment bias toward that category.

General Discussion

Across studies, participants displayed two independent social judgment biases: favoritism toward applicants who were more physically attractive and those with whom they shared an ingroup identity. Interventions that identified bias in a specific social category, indicated the direction of the bias, and asked participants to avoid showing such a bias in judgment were effective at reducing biased behavior, but only for the social categories listed in the manipulation. Unmentioned social categories showed no change in magnitude of judgment bias.

The narrow impact of the bias warning manipulation to named social categories on social judgment biases is surprising. The social categories—attractiveness, political party, university—were obvious. Faces were front and center, and university or political affiliations were indicated by a prominent image. Calling attention to potential bias by candidate attractiveness could spontaneously evoke awareness of the other obvious, irrelevant features of candidates to avoid using as a basis for judgment. Even so, unless the social category was named explicitly, the intervention had no effect. This pattern held in Study 4 when the intervention warned of different types of applicants being unfairly judged, encouraging avoidance of bias toward any irrelevant social categories.

Implications

These findings are consistent with models of bias correction suggesting that awareness of potential bias, knowledge of the direction of bias, and ability to properly adjust behavior are necessary components for reducing judgment biases (Wegener & Petty, 1997; Wilson & Brekke, 1994), and that activation of social categories and potential for bias can invoke corrective processes (Moskowitz & Li, 2011). The results add specificity to these accounts with evidence that the decision maker must be alerted to the social category (e.g., political party) rather than relying on a general appeal for avoiding bias, and that the intervention's effectiveness can be diluted by mentioning too many social categories (Study 3).

As such, this suggests that effective interventions must invoke a specific and narrow set of social categories to meet the necessary conditions for bias correction (Wilson & Brekke, 1994) or to initiate motivational processes for correcting bias to those specific social categories (Moskowitz & Li, 2011). This also suggests that alerting participants to one potential bias, and asking them to actively avoid it, may not easily provoke spontaneous introspection and generalization to other sources of bias (Wilson & Brekke, 1994), at least in the present conditions. People may need more assistance in identifying the potential sources of their biases.

Although the bias warning manipulation did not reduce favoritism for unmentioned social categories, it was able to consistently reduce bias for those social categories targeted during the intervention. Decision makers required relatively little information for the bias warning manipulations to be effective. A simple warning identifying the social group responsible for bias, specifying the direction of the bias, and prompting a commitment to avoid such biases in judgment was sufficient to reduce biased behavior (see also Axt & Nosek, 2018; Carnes et al., 2012; Pope, Price, & Wolfers, 2018). The intervention provided no concrete strategies on how to counteract bias or reminders to avoid bias. This suggests an interesting path for practical approaches to developing and testing bias intervention strategies. Bias interventions may not need to be elaborate—just specific.

In addition, while the findings reported here were robust, the effect of the bias warning manipulation for reducing targeted biases was small ($r = .095$). This relatively small effect means that follow-up research will require similarly large samples or must identify new manipulations that can produce stronger reductions in biased behavior. That said, small effects do not mean that they are inconsequential, particularly when there is reason to believe that the biases studied here have the potential to affect a wide range of important evaluations, not just selections for academic honor societies. Even small effects can produce large societal consequences when applied at scale (Greenwald, Banaji, & Nosek, 2015).

For example, a recent report noted that 106 million people use LinkedIn once a month (Yeung, 2016). A conservative estimate would be that half of these active users upload a photo to their profile (53 million). If we assume there is no correlation between attractiveness and job qualifications (Feingold, 1992), and that the people evaluating these applicants showed the same attractiveness bias as participants in our Control conditions ($r = .399$), a more physically attractive applicant would receive a more favorable evaluation 73.1% of the time when compared with an equally qualified, less physically attractive applicant. If the bias warning manipulation used here were administered to all employers using LinkedIn for interview

selection—and produced the same effect as in study participants—the percentage would drop to 67.5%. Applied to the entire pool of active LinkedIn users one time, this brief manipulation would mean that more than 1.48 million less physically attractive applicants would receive interviews that would have otherwise gone to more physically attractive applicants.

Gender as a Third Category of Social Bias

Stimuli included men and women in a fully crossed design with attractiveness and ingroup identity. Study 3 was collected after Studies 1 to 2a and Study 4, and it did not occur to us until then that we could examine gender as a third social category for which there may be a judgment bias. Indeed, in an analysis collapsing across ingroup status and physical attractiveness, there was a gender bias in control conditions such that female applicants received lower criterion than male applicants (Hedges's $g = .32$, see supplemental material for analysis from each study). This gender bias was similar in magnitude to the bias for attractiveness (Hedges's $g = .34$) and slightly larger than the bias for ingroup status (Hedges's $g = .26$).

Did the bias warning interventions about attractiveness, ingroup, or bias in general reduce this gender bias? No. Combining across Studies 1a to 3 and Study S1, warning about biases or prejudice for one or two of the other social categories did not affect the gender bias (Hedges's $g = –.01$), and warning about bias in general in Study 4 also did not alter gender bias ($d = −0.04$). These exploratory results are consistent with the confirmatory findings that drawing attention to a bias in a specific social category is a necessary component for bias reduction. However, we do not have evidence that drawing attention to gender in a more focused intervention would reduce gender bias. It is conceivable that gender biases are not amenable to similar interventions. Nevertheless, these exploratory results are consistent with the conclusion that warning about bias in a specific social category is necessary for bias reduction, and the lack of reduction of gender bias in Study 3 affirms that there is an upper limit to naming social categories for an intervention to retain its effectiveness.

Moderation by Desired Performance

Primary analyses focused on the impact of the bias warning interventions on all participants, regardless of their desired performance, and results found that warnings about a specific bias were associated with reduced criterion biases on that dimension. However, it is possible that the impact of such interventions is limited to participants who reported a desire to act fairly, and that the bias warning interventions are ineffective on participants who report a desire to favor more physically attractive applicants or applicants from one's ingroup.

To address this question, we ran another exploratory analysis. For each study, participants' desired responses were recoded into either wanting to show no bias or wanting to favor either more attractive people or ingroup members. Unsurprisingly, bias was large among participants who wanted to show favoritism (Hedges's $g = .67$, $p < .001$). Critically, bias was also observed among control participants who wanted to be fair (Hedges's $g = .23$, $p < .001$). Among participants who reported wanting to be fair, bias warning manipulations mentioning a specific bias had a small but reliable effect (Hedges's $g = .14$, $p < .001$), indicating a decrease in criterion bias following a bias warning manipulation. And, even among participants with a desire to show favoritism, there was evidence that the bias warning manipulation also reduced criterion bias (Hedges's $g = .22$, $p = .005$).

The bias warning manipulation appeared to reduce bias among those both wanting and not wanting to show favoritism, and it is possible that the psychological mechanism behind the manipulation's effectiveness is different for these two groups of participants. Among participants with a desire to behave fairly, the bias warning manipulation may have guided attention to the impact of a social dimension that would have otherwise gone unnoticed or was noticed but not taken seriously as a source of bias. Among participants with a desire to show bias, the warning manipulation may have simply weakened that desire (as suggested by the meta-analysis reported in Study 1a), which in turn resulted in less bias on the JBT. Or, the manipulation may have created a distinction between how participants *wanted* to behave and how they *should* behave; although they wanted to favor some groups over others, the intervention may have increased recognition that it was inappropriate to do so. Future studies could clarify the role of motivation in bias-reduction interventions.

Changing Bias Versus Ability to Show Unbiased Behavior

The manipulations used here alerted participants to one or two specific social dimensions known to bias judgment and asked them to refrain from using that information when evaluating applicants from an honor society. Being warned about a specific bias in evaluation reduced judgment biases concerning that specific social dimension. However, it is unlikely that the bias warning intervention was effective because it lessened participants' attitudinal preferences, which then led to reduced judgment biases. Our meta-analysis showed that the manipulation created very small reductions in explicit preferences (Hedges's $g = .05$), but participants in bias warning conditions still reported large preferences for physically attractive people (aggregate $d = 0.65$) and ingroup members (aggregate $d = 1.10$).

It is more likely that the bias warning manipulation was effective because it either increased participants' ability to control their biases or their motivation to do so (Fazio, 1990; Wilson & Brekke, 1994). The domains in these studies were ones where participants felt free to report explicit preferences, and also recognized that it would be inappropriate to use those preferences to guide judgments in an academic context. Among participants who reported an explicit preference for more physically attractive people, 88% reported

wanting to treat more and less physically attractive people equally on the JBT, and 81% of participants reporting an explicit preference for political or university ingroup members reported wanting to not use group membership when evaluating applicants.

Constraints on Generalizability and Next Steps

Our conclusion is characterized in general terms—to reduce bias toward members of that social category, it is necessary to explicitly identify the social category. The actual generalizability of the present evidence for this conclusion is unknown. The constraints on that generalizability will be advanced by replicating this investigation with a variety of methodological changes.

First, it is possible that this conclusion is constrained to idiosyncratic features of the selected social categories (attractiveness, institution or political affiliation, gender) or to interactions between these categories. For example, biases toward the attractiveness and ingroup categories were very weakly correlated (aggregate $r = .04$), perhaps reducing the likelihood that an intervention targeting one would affect the other. If the biases were more strongly correlated, then it is easy to assume that an intervention targeting one would be more likely to generalize to others. Of interest would be whether some correlation between biases is necessary to invoke the spontaneous generalization that would allow for an intervention targeting one social category to be equally effective at reducing bias concerning another social category.

Similarly, some social categories may be particularly prone to spontaneous awareness whenever the potential for bias is mentioned, meaning certain interventions may be effective for salient categories without needing to explicitly identify the social category. Across studies, it was unambiguous that most participants did not intend to favor applicants based on their physical attractiveness or their political affiliation; 91.8% of participants reported not wanting to use physical attractiveness and 86.3% reported not wanting to use political affiliation when evaluating applicants. Yet, while these attractiveness and political biases were unintended, it is unclear the extent to which participants saw them as problematic (i.e., how bothered they would have been if presented with evidence that their judgments were indeed biased on these social dimensions). It is possible that the bias warning manipulation could have produced stronger effects or effects that carried over into unmentioned social categories, if the task focused on more socially sensitive dimensions, such as sexual orientation. In addition, it is conceivable that more general bias warnings, like those used in Study 4, would be effective at reducing biases along other social dimensions, such as race.

Second, our conclusion may be constrained to the features of the judgment bias paradigm—the JBT, the outcomes for which the applicants were being judged, or the criteria on which they were judged. For example, similar effects might not emerge with budgeting allocations (Rudman & Ashmore, 2007). We have no theoretical reason to expect that effects are contingent on this paradigm, but we do not have evidence affirming or discounting its generalizability across such features. Similarly, it is possible that our conclusions are limited to this version of the JBT. The JBT is, by design, a relatively challenging task with objectively correct answers, and it is highly plausible that biases are easier to control and avoid when tasks are simple and correct answers are clear. With simpler versions of the task, a bias warning intervention may be more likely to generalize across multiple social categories because the performance criteria are unambiguous. Similarly, with simpler tasks, participants' intentions, attitudes, and self-assessments of performance might be more strongly correlated with their actual task performance because it is easier to apply one's intentions on actual performance.

Third, it is possible that this conclusion is constrained to features of the intervention. For example, it is possible that other interventions do not require warning about the specific social category to reduce bias or do, in fact, generalize across biases. Such interventions could target specific biases influencing judgment, such as imagining contact with outgroup members (Todd, Bodenhausen, Richeson, & Galinsky, 2011), or could instill a more general mind-set capable of affecting reliance on all irrelevant social information, such as through self-distancing (Ayduk & Kross, 2010). Despite a wealth of research using different bias reduction strategies, few have directly manipulated mentioning the social category of interest to evaluate whether it is necessary for intervention effectiveness, and none have tested whether the interventions generalize to unmentioned social categories.

It is also possible that different versions of the bias warning manipulation could have created effects that affected both mentioned and unmentioned biases. We tested one possibility based on a reviewer's suggestion: that using the more affectively laden word "prejudiced" in the manipulation could produce differing results than the word "biased." We ran a replication of Study 2a (Study S1 $N = 1{,}540$), with the only change being that the word "biased" was replaced with the word "prejudiced" in the manipulation text (see https://osf.io/ms9ye/ for the study's preregistration). Results replicated those of Studies 2a and 2b. Participants warned about a specific bias (or prejudice) showed reduced criterion bias for that social category (all η_p^2s$>.011$; all ps $< .001$), but there was no reliable evidence that being warned about a specific bias (or prejudice) reduced criterion bias for the other, unmentioned social category (all η_p^2s$<.001$; all ps $> .364$; see supplemental material for full reporting). Regardless, it is possible that future versions of the manipulation will produce changes in biased behavior for both targeted and untargeted biases.

Fourth, it is possible that our conclusion is constrained to samples that we investigated. Most of this research was conducted with a heterogeneous cross section of adults visiting a public website. It is notable that so many biases were observed considering the website itself is associated with investigating bias. To the extent that participants were aware of that association, the bias effects may be smaller than

would be observed in other circumstances. Nevertheless, reliable biases were observed. An open question is whether the intervention strategies will be any more or less effective in other sampling contexts.

Finally, our bias warning intervention occurred immediately prior to evaluation. We have no evidence about what time or distractions between intervention and evaluation might have on the interventions' effectiveness. Other research shows that related interventions have short-lived effects (e.g., Lai et al., 2016). There must be boundary conditions on time and other interference between the intervention and the judgment, but we have no evidence yet to calibrate such constraints on generalizability.

Conclusion

An intervention warning about the potential for bias reduced social biases in judgment, but only when alerting participants to the specific social category affecting evaluation. The intervention did not affect other social biases affecting judgment. This implies that identifying the social category responsible for bias is necessary for reducing bias toward that category. If this conclusion is generalizable across a variety of social categories, intervention types, and outcomes, then the implications would be quite substantial for theoretically understanding how social judgment biases can be reduced and practically for developing intervention strategies. The next stage of research should identify conditions under which our conclusion does not hold.

Authors' Note

Studies 1a and 1b were originally from the first author's dissertation.

Acknowledgments

We thank Martha Fulp-Eickstaedt, Natalie Buchen, and Anran Xiao for assistance in data collection.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: This research was partly supported by Project Implicit. B.A.N. is an officer and J.R.A. is Director of Data and Methodology for Project Implicit, Inc., a nonprofit organization with the mission to "develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender, or other factors."

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Study 1a was partly supported by an Ingrassia Family Echols Scholars Research Grant awarded to the second author.

Note

1. The increased exclusion rate is likely due to the lack of encoding phase, which another study found decreased initial dropout but led to higher exclusion rates (Axt, Nguyen, & Nosek, 2018). None of the primary conclusions change when including all participants (see supplemental material).

Supplemental Material

Supplemental material is available online with this article.

References

Axt, J.R. (2017). *The impact of awareness on reducing social bias in behavior*. (Unpublished doctoral dissertation). University of Virginia, Charlottesville, VA.

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2016). An unintentional, robust, and replicable pro-Black bias in social judgment. *Social Cognition*, *34*(1), 1-39.

Axt, J. R., Nguyen, H., & Nosek, B. A. (2018). The judgment bias task: A reliable, flexible method for assessing individual differences in social judgment biases. *Journal of Experimental Social Psychology*, *76*, 337-355.

Axt, J. R., & Nosek, B. A. (2018). Awareness may be the mechanism for multiple interventions that reduce social judgment biases. Manuscript submitted for publication.

Ayduk, Ö., & Kross, E. (2010). From a distance: Implications of spontaneous self-distancing for adaptive self-reflection. *Journal of Personality and Social Psychology*, *98*, 809-829.

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

Carnes, M., Devine, P. G., Isaac, C., Manwell, L. B., Ford, C. E., Byars-Winston, A., . . . Sheridan, J. (2012). Promoting institutional change through bias literacy. *Journal of Diversity in Higher Education*, *5*(2), 63-76.

Cole, E. R. (2009). Intersectionality and research in psychology. *American Psychologist*, *64*, 170-180.

Derous, E., Ryan, A. M., & Serlie, A. W. (2015). Double jeopardy upon resume screening: When Achmed is less employable than Aisha. *Personnel Psychology*, *68*, 659-696.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*, 1267-1278.

Devine, P. G., Forscher, P. S., Cox, W. T., Kaatz, A., Sheridan, J., & Carnes, M. (2017). A gender bias habit-breaking intervention led to increased hiring of female faculty in STEMM departments. *Journal of Experimental Social Psychology*, *73*, 211-215.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75-107). New York, NY: Academic Press.

Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin*, *111*, 304-341.

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology*, *72*, 133-146.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*, 535-549.

Golding, J. M., Fowler, S. B., Long, D. L., & Latta, H. (1990). Instructions to disregard potentially useful information: The

effects of pragmatics on evaluative judgments and recall. *Journal of Memory and Language*, *29*, 212-227.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*, 493-503.

Greenman, E., & Xie, Y. (2008). Double jeopardy? The interaction of gender and race on earnings in the United States. *Social Forces*, *86*, 1217-1244.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*, 553-561.

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*, 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197-216.

Hawkins, C. B., & Nosek, B. A. (2012). Motivated independence? Implicit party identity predicts political judgments among self-proclaimed independents. *Personality and Social Psychology Bulletin*, *38*, 1437-1452.

Hosoda, M., & Stone-Romero, E. (2010). The effects of foreign accents on employment-related decisions. *Journal of Managerial Psychology*, *25*, 113-132.

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology*, *41*, 137-146.

Kang, S. K., & Bodenhausen, G. V. (2015). Multiple identities in social perception and interaction: Challenges and opportunities. *Annual Review of Psychology*, *66*, 547-574.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., . . . Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*, 1765-1785.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., . . . Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*, 1001-1016.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*, 255-275.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology*, *83*, 1029-1050.

Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, *77*, 167-184.

Moskowitz, G. B., & Li, P. (2011). Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *Journal of Experimental Social Psychology*, *47*, 103-116.

Moskowitz, G. B., Salomon, A. R., & Taylor, C. M. (2000). Preconsciously controlling stereotyping: Implicitly activated egalitarian goals prevent the activation of stereotypes. *Social Cognition*, *18*, 151-177.

Mullen, B., Brown, R., & Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology*, *22*, 103-122.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265-292). New York, NY: Psychology Press.

Pingitore, R., Dugoni, B. L., Tindale, R. S., & Spring, B. (1994). Bias against overweight job applicants in a simulated employment interview. *Journal of Applied Psychology*, *79*, 909-917.

Pope, D. G., Price, J., & Wolfers, J. (2013). *Awareness reduces racial bias* (NBER Working Paper 19765). Retrieved from http://www.nber.org/papers/w19765

Pope, D.G., Price, J. & Wolfers, J. (2018). Awareness reduces racial bias. *Management Science.* Advance online publication. Retrieved from https://pubsonline.informs.org/doi/pdf/10.1287/mnsc.2017.2901.

Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the Implicit Association Test. *Group Processes & Intergroup Relations*, *10*, 359-372.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*, 293-304.

Sassenberg, K., & Moskowitz, G. B. (2005). Don't stereotype, think different! Overcoming automatic stereotype activation by mindset priming. *Journal of Experimental Social Psychology*, *41*, 506-514.

Schul, Y. (1993). When warning succeeds: The effect of warning on success in ignoring invalid information. *Journal of Experimental Social Psychology*, *29*, 42-62.

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology*, *56*, 283-294.

Todd, A. R., Bodenhausen, G. V., Richeson, J. A., & Galinsky, A. D. (2011). Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology*, *100*, 1027-1042.

Wegener, D. T., & Petty, R. E. (1997). The flexible correction model: The role of naive theories of bias in bias correction. *Advances in Experimental Social Psychology*, *29*, 141-208.

Wegner, D. M., Coulton, G. F., & Wenzlaff, R. (1985). The transparency of denial: Briefing in the debriefing paradigm. *Journal of Personality and Social Psychology*, *49*, 338-346.

Wetzel, C. G., Wilson, T. D., & Kort, J. (1981). The halo effect revisited: Forewarned is not forearmed. *Journal of Experimental Social Psychology*, *17*, 427-439.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*, 117-142.

Woodhams, C., Lupton, B., & Cowling, M. (2015). The snowballing penalty effect: Multiple disadvantage and pay. *British Journal of Management*, *26*, 63-77.

Yeung, K. (2016, August 4). *LinkedIn now has 450 million members, but the number of monthly visitors is still flat*. Available from http://www.venturebeat.com