Manuscript Under Review

Reducing social judgment biases may require identifying the potential source of bias

Jordan R. Axt

University of Virginia

Project Implicit

Grace Casola

University of Virginia

Brian A. Nosek

University of Virginia

Center for Open Science

Word Count (including text, abstract, notes and references): 9,975

Author Contact Information
Jordan Axt
University of Virginia
Box 400893
Charlottesville VA, 22904-4400
jra3ee@virginia.edu

Acknowledgments: We thank Martha Fulp-Eickstaedt, Natalie Buchen, and Anran Xiao for assistance in data collection. Study 1a was partly supported by an Ingrassia Family Echols Scholars Research Grant awarded to the second author.

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

Declaration of Conflicting Interests: This research was partly supported by Project Implicit. B. A. Nosek is an officer and J.R. Axt is a research associate 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." The authors declare no other potential conflicts of interest with respect to authorship or publication of this article.

Abstract

Social judgment is shaped by multiple biases operating simultaneously, but most bias-reduction interventions target only a single social category. In four pre-registered studies (Total N > 4,800), we investigated whether raising awareness of one social bias impacted that and other social biases. Participants selected honor society applicants based on academic credentials. Applicants also differed on social categories that were irrelevant for selection: attractiveness and ingroup status. Participants alerted to potential bias in one social category showed reduced bias for that category (average η^2_p = .01), but showed no bias reduction on the unmentioned social category (average η^2_p = .002). In a fifth study, raising awareness of social biases without specifically identifying a category did not reduce bias on either social category (average η^2_p = .002). These results suggest that effectiveness of raising awareness on reducing social biases is highly specific, perhaps even contingent on explicitly identifying the potential source of bias.

Word count: 148

Keywords: Bias, social judgment, awareness, discrimination, prejudice

Reducing social judgment biases may require identifying the potential source of bias

Whether 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 non-overweight candidates (Pingitore, Dugoni, Tindale & Spring, 1994). Biases may also emerge when people use group identity to inform evaluation, such as when foreign applicants are rates as less hirable than native applicants (Hosoda & Stone-Romero, 2010).

There are many possible interventions that could reduce such biases (e.g., Gollwitzer, 1999; Lerner & Tetlock, 1999). One of the most direct is to make people aware of potential bias in judgment, so that 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, 1996; Wilson & Brekke, 1994). Awareness interventions 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 employ a paradigm for which an awareness intervention reliably reduces social judgment biases (Axt, 2017).

Most research on social judgment biases examines a single category at a time such as race, age, or weight, presumably for experimental simplicity. But, in reality, people are all those things at once. 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 awareness interventions impact 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" information from other irrelevant categories. In the present research, we investigated whether an awareness intervention had a constrained impact to the social categories identified in the intervention itself, or a general impact on social categories whether they were explicitly identified in the intervention or not.

Existing Evidence

Research on multiple identities and intersectionality has examined how multiple social categories independently or interactively influence evaluation of a single target (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, 2013). Similar results emerged in an analysis of payment towards employees that were both racial and gender minorities (Greenman & Xie, 2008). Field experiments have found that recruiters rated applicants with multiple stigmatized identities (e.g., ethnic and gender identities associated with greater threat) as less employable than applicants with a single threat-related identity (Derous, Ryan & Serlie, 2015).

These examples show a healthy literature examining 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 toward social groups (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: (1) awareness of unwanted processing related to the judgment, (2) motivation to correct bias, (3) awareness of direction and magnitude of bias, and (4) ability to adjust responses. A failure to meet any condition results in biased judgment. If decision-makers are not aware of the unwanted bias, then there is no opportunity for motivation and corrective processes to engage and reduce bias.

For situations in which two social categories simultaneously contaminate evaluation of individual targets, the model would most directly predict that raising awareness of potential 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, raising awareness of a bias concerning one social category might lead people to spontaneously invoke awareness of potential biases concerning other social categories. Moreover, the model is not definitive on specificity of awareness. For instance, for reducing political orientation bias, would warning people of bias toward “people from different categories” be as effective as warning of 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” 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 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 created increased concern about racial discrimination (Devine et al., 2012; Forscher et al., 2017).

These motivational accounts are also non-specific as to whether awareness of one bias would create reductions in other simultaneous biases. On the 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 findings that priming a general mindset 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 they share an expectation that activation or awareness of potential bias is important for changing bias in behavior, and they share a lack of specificity for whether awareness of potential bias for one social category is sufficient to reduce bias toward another social category. As such, both models of social judgment biases will be improved by the evaluating whether awareness of the particular social category is necessary for bias reduction for that category.

The Present Work

In all studies, participants completed a Judgment Bias Task (JBT; Axt, Nguyen & Nosek, 2017). In the JBT, participants evaluate profiles for an outcome, here admission to an honor society. Each profile includes quantified criteria relevant for evaluation (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 impacts 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: a face high or low 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. 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 raised awareness of either one or both of the physical attractiveness or ingroup biases. We investigated whether the awareness interventions reduced bias for the specific category only, or whether it also reduced the unmentioned category bias. The evidence indicates that awareness interventions are effective at reducing biases for the identified category but not for others. In Study 3, we tested whether mentioning 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.

Study 1a-1b

In Study 1a, profiles were presented with more or less physically attractive faces and images indicating applicants came from participants’ 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 possibly showing ingroup favoritism.

Method

Participants

Participants in Study 1a were University of Virginia 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* = .22, which was the size of the smaller criterion bias found in a pilot study (see online supplement). 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*-augmented (Sagarin, Ambler & Lee, 2014). The final sample had 929 participants (M_{Age} = 18.9, *SD* = 1.2, 62.5% female, 59.4% White). See https://osf.io/bm5yk/ for pre-registration 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 1200 participants, which would provide greater than 93% power at detecting at least suggestive evidence ($p < .05$; Benjamin et al., 2017) for a between-subjects effect of *d* = .20. In the final sample, 1223 participants provided data and passed the attention check[1] (M_{Age} = 42.4, *SD* = 13.0, 72.6% female, 63.3% White). See https://osf.io/2dpmx/ for the study's pre-registration.

[1] Conclusions do not change when including participants who did not pass the attention check (*N* = 1157; see online supplement).

We report all measures, manipulations, and exclusion criteria. Materials, data and online supplements are available at https://osf.io/mqdga/.

Procedure

Participants in Study 1b completed four study components in the following order: participants first received the awareness 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 Awareness condition. Participants in the Control condition received no additional instructions. In Study 1a, participants in the Awareness condition were alerted to a bias favoring applicants from one's own university. In Study 1b, participants in the Awareness condition were alerted to a bias 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 Awareness manipulation was the same for Study 1a, except "university" replaced "political party".

Existing evidence shows that this awareness manipulation reduced bias in criterion on a JBT (Axt, 2017). The manipulation also asks participants to be fair when evaluating applicants, which could influence its effectiveness. However, in prior research, a version of the intervention without the fairness text was no less effective than a version including both awareness and fairness components (Axt, 2017, Study 2). Moreover, an intervention emphasizing “be fair” without awareness was not effective (Axt, 2017, Study 3). Given this prior work, our discussion of the intervention places emphasis on its manipulating awareness of a potential bias.

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 1-4 scale. The two GPAs already had a maximum score of four. Recommendation letters were scored Poor = 1, Fair = 2, Good = 3, 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 pre-tested to differ and divided into two groups differing in physical attractiveness (*d* = 2.64; Axt, Nguyen & Nosek, 2017). In Study 1a, applicants were also depicted with a logo from the University of Virginia (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 lasted for 64 trials, with eight trials (four male, four female) 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.). Each applicant had a unique combination of academic qualifications. Before evaluating applicants, participants completed an encoding phase where each applicant was shown for one second in a random order, though this encoding phase was removed from Study 1b to save time. Evaluations were made one at a time 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 versus less physically attractive face and to be depicted as from UVA vs. 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.

Awareness of 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 seven-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 seven-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 seven-block evaluative Implicit Association Tests (IAT; Greenwald, McGhee & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2007) measuring evaluations towards more vs. less physically attractive people and identification with UVA vs. UNC. See the online supplement for information on IAT procedure. IATs were completed in a random order and scored with the *D* algorithm (Greenwald, Nosek & Banaji, 2003).

Analyses concerning changes in desired performance, perceived performance and attitudes, or how these variables relate to JBT performance, were not the central focus of this investigation. As a result, we placed all additional pre-registered or exploratory analyses regarding these issues in the supplementary materials available at https://osf.io/8rkfj/. We observed effects consistent with findings reported in Axt, Ebersole & Nosek, 2016 and Axt, Nguyen & Nosek, 2017).[2]

[2] Across studies, correlations with JBT performance consistently replicate the findings of prior work (Axt, Ebersole & Nosek, 2016; Axt, Nguyen & Nosek, 2017): criterion bias was weakly

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-3, we defined Democrats and Republicans as those selecting that party identification immediately and those that selecting that party in the forced choice. Other evidence suggests these groups are similar in political judgment (Hawkins & Nosek, 2012), and combining them maximized 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

but positively associated with explicit attitudes (*r* = .18, 95% CI [.12, .23]), perceived performance (*r* = .28, 95% CI [.19, .38]), desired performance (*r* = .21, 95% CI [.12, .30]), and implicit associations (*r* = .12, 95% CI [.09, .15]). We also tested whether awareness of a specific bias was associated with attitudes or the performance measures. Across studies, we found no effects that awareness of a specific bias changed implicit associations (Hedge's *g* = .01, 95% CI = [-.11, .12]), but did find small effects that awareness reduced explicit preferences (Hedge's *g* = .07, 95% CI = [.03, .12]). And, in line with actual JBT behavior, awareness was associated with small changes in desired performance (Hedge's *g* = .12, 95% C.I. [.04, .19]) and perceived performance (Hedge's *g* = .19, 95% CI = [.13, .26]) Perceived and desired performance indicated more equal treatment following the awareness intervention. See online supplement for results from individual studies.

(Axt, Ebersole & Nosek, 2016; Axt, Nguyen & Nosek, 2017). Fifteen participants (1.6%) were excluded based on these criteria in Study 1a and 286 participants (24.1%) in Study 1b.[3]

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: Awareness vs. Control) mixed-measures ANOVA on criterion for applicants from each combination of school and physical attractiveness. This analysis revealed main effects of physical attractiveness, $F(1, 912) = 130.58$, $p < .001$, $\eta^2_p = .125$, 95% C.I. [.09, .17], and school, $F(1, 912) = 8.71$ $p = .003$, $\eta^2_p = .009$, 95% C.I. [.001, .03], *p*-augmented = [.05, .0502], with lower criterion for more vs. less physically attractive applicants and for applicants from UVA vs. UNC. There was no evidence of a main effect of Condition, $F(1, 912) = .42$, $p = .516$, $\eta^2_p < .001$, 95% C.I. [0, .01]. See Figure 1 for criterion values in each condition and combination of school affiliation and attractiveness, and Table 1 for means and standard deviations in each condition for all studies.

Of primary interest were the attractiveness by condition and school by condition interactions. A school by condition interaction could indicate that awareness of a potential school-affiliation bias reduced the bias favoring applicants from one's own university. The school by condition interaction pattern was consistent with this prediction but did not provide

[3] 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, 2017). None of the primary conclusions change when including all participants (see online supplement).

strong evidence, *F*(1, 912) = 3.03, *p* = .082, η2p = .003, 95% C.I. [0, .02], *p*-augmented = [.082, .119]. The main effect of school was larger in the Control (*F*(1, 424) = 9.25 *p* = .002, η^2_p = .021, 95% C.I. [.003, .06], $p_{\text{augmented}}$ = [.05, .0501]) than the Awareness (*F*(1, 488) = .87, *p* = .351, η^2_p = .002) condition.

A reliable attractiveness by condition interaction could indicate that awareness of a school-affiliation bias reduced the criterion bias favoring more physically attractive applicants, which would suggest that awareness of a specific bias reduced other biases. The attractiveness by condition interaction was small but suggestive, *F*(1, 912) = 5.13, *p* = .024, η2p = .006, 95% C.I. [0, .02], *p*-augmented = [.051, .058]. The main effect of attractiveness was larger in the Control (F(1, 424) = 80.37, *p* < .001, η2p = .159, 95% C.I. [.10, .22]) than the Awareness (F(1, 488) = 48.96, *p* < .001, η2p = .091, 95% C.I. [.05, .14]) condition.

Unrelated to key hypotheses, neither the school by attractiveness interaction, F(1, 912) = 3.72, *p* = .054, η2p = .004, 95% C.I. [0, .016], or the school by attractiveness by condition interaction, F(1, 912) = 2.23, *p* = .136, η2p = .002, 95% C.I. [0, .013], produced strong effects. Study 1a results were consistent with the hypothesis that awareness 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: Awareness vs. Control) mixed-measures measures ANOVA on criterion for applicants from each combination of political party and physical attractiveness. This analysis revealed main effects of physical attractiveness, *F*(1, 898) = 105.77, *p* < .001, η2p = .105, 95% C.I. [.07, .14], and political ingroup, *F*(1, 898) = 65.82 p < .001, η2p = .068, 95% C.I. [.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, $p = .441$, η2p = .001, 95% C.I. [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, $p = .001$, η2p = .013, 95% C.I. [.003, .03]. The main effect of school was larger in the Control (*F*(1, 466) = 55.78, $p < .001$, η2p = .107, 95% C.I. [.06, .16]) than the Awareness (*F*(1, 432) = 14.18, $p < .001$, η2p = .032, 95% C.I. [.01, 07]) condition. Making participants aware of 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, $p = .336$, η2p = .001, 95% C.I. [0, .009]. The main effect of attractiveness in the Control condition (*F*(1, 466) = 44.65, $p < .001$, η2p = .087, 95% C.I. [.04, .14]) was, if anything, slightly smaller than in the Awareness condition (*F*(1, 432) = 61.70, p < .001, η2p = .125, 95% C.I. [.07, .18]). Making participants aware of potential political bias did not appear to reduce expression of attractiveness bias.

There was not strong evidence for either a political group by attractiveness interaction, F(1, 898) = .10, p = .755, η2p < .001, nor a political group by attractiveness by condition interaction, F(1, 898) = .97, p = .325, η2p = .001. Study 1b suggested that awareness of potential bias for one social category reduced bias for that category but not another social category.

Discussion

In control conditions, participants displayed two simultaneous biases: favoritism toward more physically attractive people and toward ingroup members. Making participants aware of potential university or political bias reduced that bias, though very weakly in Study 1a. And, making participants aware of potential 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 awareness of one category bias influences a second, unmentioned social category bias. We sought more evidence by replicating and extending these findings. In Studies 2a-2b, we replicated Study 1b and extended it into a 2x2 design making participants aware of none, one, or both of the bias categories. This design would provide more comprehensive evidence about whether it is necessary to identify the social category for interventions aimed at reducing social biases to be effective.

Study 2a-2b

Method

Participants

Participants in Studies 2a and 2b came from the Project Implicit research pool. Participants were selected from a pool of available studies if they met selection criteria. In Study 2a, we selected Americans who reported being at least slightly liberal or conservative, and targeted a sample of 400 in each condition. This sample would provide over 90% power at detecting a between-subjects effect of *d* = .24, which was the size of the impact of awareness on 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/9f83w/ for the study’s pre-registration.

In Study 2b, we selected Americans that 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* = .21, which was the average effect for the impact of raising awareness of a specific bias on the in Studies 1b and 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/zg5ut/ for the study's pre-registration.

Procedure

In both studies, participants completed five components in the following order: awareness 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 awareness of differences among applicants' attractiveness, gender, race, political party, and GPA. Our pre-registered 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 Awareness, Attractiveness Awareness, or Dual Awareness. In the Control condition, participants received no additional instructions. In the Politics Awareness condition, participants read the same manipulation as Study 1b. In the Attractiveness Awareness condition, participants read an updated version of that manipulation replacing any mention of political ingroup bias with a bias favoring more physically attractive people. In the Dual Awareness condition, participants read the relevant text from both the Attractiveness and Politics Awareness conditions. See online supplement 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 vs. Republicans.

Results

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 Awareness Condition: Awareness vs. None) by 2 (Politics Awareness Condition: Awareness vs. None) mixed-measures ANOVA on criterion for each combination of applicant attractiveness and political identity.

Both studies showed main effects of physical attractiveness (Study 2a: *F*(1, 1862) = 47.58, *p* < .001, η2p = .025, 95% C.I. [.013, .041]; Study 2b: *F*(1, 1515) = 92.90, *p* < .001, η2p = .058, 95% C.I. [.037, .082]), such that more physically attractive applicants received lower criterion than less physically attractive applicants. There were also main effects of applicant political party (Study 2a: *F*(1, 1862) = 57.61, *p* < .001, η2p = .030, 95% C.I. [.017, .047]; Study 2b: *F*(1, 1515) = 56.10, *p* < .001, η2p = .036, 95% C.I. [.020, .056]), such that applicants from one's own party received lower criterion than from the other party. Main effects of attractiveness or politics awareness condition were not reliable (η2p < .002). See Figure 3 for criterion values in Study 2a, Figure 4 for Study 2b.

If raising awareness of a specific bias reduced that bias, there would be evidence of interactions between applicant attractiveness and attractiveness awareness condition and between applicant political ingroup and politics awareness condition. Both studies showed an applicant attractiveness by attractiveness awareness interaction (Study 2a: $F(1, 1862) = 35.60$, $p < .001$, η2p = .019, 95% C.I. [.009, .033]; Study 2b: $F(1, 1515) = 9.40$, $p = .002$, η2p = .006, 95% C.I. [.001, .016]); the attractiveness bias was smaller when participants were made aware of that potential bias (Study 2a: η2p = .0004; Study 2b: η2p = .027) than when not (Study 2a: η2p = .086; Study 2b: η2p = .100).

There was weak evidence of political ingroup by politics awareness interactions in Study 2a, $F(1, 1862) = 5.19$, $p = .023$, η2p = .003, 95% C.I. [.00002, .010], and Study 2b, $F(1, 1515) = 3.34$, $p = .068$, η2p = .002, 95% C.I. [0, .009]. In both cases, the main effect of political ingroup was smaller when participants had been made aware of the bias (Study 2a: η2p = .018; Study 2b: η2p = .024) than when not (Study 2a: η2p = .043; Study 2b: η2p = .047). The evidence for awareness reducing bias was consistent for attractiveness and politics, but stronger for attractiveness.

A central question was whether raising awareness of bias for one category influenced the bias expressed on the other category. If so, we would observe interactions between applicant attractiveness and politics awareness conditions and between applicant political ingroup status and attractiveness awareness conditions. In both studies, there was no evidence of interactions between applicant attractiveness and politics awareness condition (Study 2a: $F(1, 1862) = .03$, p

= .868, η2p < .001; Study 2b: *F*(1, 1515) = .03, *p* = .865, η2p < .001), or between applicant political ingroup by attractiveness awareness condition (Study 2a: *F*(1, 1862) = .94, *p* = .333, η2p = .001; Study 2b: *F*(1, 1515) = .11, *p* = .744, η2p < .001).

It is possible that the effectiveness of awareness interventions was moderated by whether participants were also made aware of the other potential source of bias. This would be indicated by 3-way interactions of both awareness manipulations and the category manipulation (attractiveness or politics). There was no evidence of an interaction between applicant attractiveness and the two awareness conditions (Study 2a: *F*(1, 1862) = .29, *p* = .593, η2p < .001; Study 2b: *F*(1, 1515) = .13, *p* = .720, η2p < .001). Also, there was no evidence of an interaction between applicant politics, attractiveness awareness, and politics awareness in Study 2b (*F*(1, 1515) = .46, *p* = .500, η2p < .001). However, there was suggestive evidence of such an interaction in Study 2a (*F*(1, 1862) = 7.37, *p* = .007, η2p = .004, 95% C.I. [.0003, .012]). The impact of the politics awareness manipulation on reducing political bias was stronger when participants were only made aware of the politics bias (η2p = .012) compared to when they were also aware of an attractiveness bias (η2p < .001). While intriguing, as a single instance among multiple tests, this is weak evidence that should be replicated before taking seriously.

No other terms in the ANOVA in either study reached suggestive (*p* < .05) or stronger (*p* < .005) statistical significance (Benjamin et al., 2017). See online supplement for full reporting.

Discussion

Making participants aware of 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 awareness of potential bias toward one social category reduced bias toward another social category.

Internal Meta-Analysis of Awareness

The influence of the awareness manipulation across Studies 1a-2b yielded somewhat variable results. In Study 1a, there were very small effects of awareness reducing the targeted bias (favoring of one's own university) and the untargeted bias (favoring more physically attractive people). In Studies 1b-2b, there was no effect of awareness reducing the untargeted bias and only an impact of awareness reducing the targeted bias, though this effect was quite weak with awareness of political bias in Study 2b.

To provide more precise estimates of the impact of awareness on targeted and untargeted biases, we conducted a "mini meta-analysis" (Goh, Hall & Rosenthal, 2016) of Studies 1a-2b, converting each partial eta-squared (η2p) into a Pearson's correlation (*r*). Higher *r* values mean the awareness intervention more effectively reduced criterion bias. There was a small but reliable meta-analytic effect that raising awareness of a bias concerning a social category was associated with reduced criterion bias toward that category (*r* = .082, 95% C.I [.050, .114], Z = 5.07, *p* < .001). However, there was no reliable meta-analytic effect of raising awareness of bias concerning one social category reducing criterion bias toward the unmentioned social category (*r* = .007, 95% C.I [-.015, .028], Z = 0.58, *p* = .557). See Figure 5 for forest plots. The meta-analytic results provide a relatively unambiguous conclusion--interventions raising awareness and encouraging fairness toward one social category were effective at reducing social judgment bias toward that category, but not toward another, unmentioned social category.

Study 3

Studies 1a-2b suggest that awareness of the social category may be a necessary condition for bias-reduction interventions to be effective. This possibility introduces an obvious question --what happens with the same intervention, but instead of highlighting any specific social category the intervention refers 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 3, we conduct a high-powered test of this possibility.

Method

Participants

Using the Project Implicit research pool, 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 1-2b for the impact of raising awareness of a specific bias (d = .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/7mwpf/ for the study’s pre-registration.

Procedure

Participants completed the same measures in the same order as Study 2a, except for the contents of the awareness manipulation.

Experimental conditions. Participants were randomly assigned to a Control or General Awareness condition. In the Control condition, participants received no additional instructions. In the General Awareness condition, participants received the same manipulation as participants in the Politics Awareness conditions as Studies 2a-2b, except any mention of the politics social category was removed. Specifically, participants read:

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.*

Participants then wrote that they would be fair in a text box.

JBT and follow-up measures. Participants completed the same JBT, measures of performance and explicit attitudes, and measure of implicit political identification as in Study 2a.

Results

85 (6.8%) participants were excluded from analysis. Accuracy on the task 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 Awareness) mixed-measures ANOVA on criterion. This analysis revealed main effects of physical attractiveness, *F*(1, 1162) = 159.29, *p* < .001, η2p = .120, 95% C.I. [.09, .16], and political ingroup, *F*(1, 1162) = 56.60, *p* < .001, η2p = .046, 95% C.I. [.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) < .001, *p* = .996, η2p < .001.

Of primary interest were the attractiveness by awareness and politics by awareness interactions, which could indicate that the general awareness manipulation impacted evaluation of more vs. less physically attractive applicants or applicants from one's own vs. the other political party. There was no evidence for interactions between applicant attractiveness and awareness condition, (*F*(1, 1162) = 1.74, *p* = .188, η2p = .001), or between applicant political ingroup and awareness condition (*F*(1, 1162) = .32, *p* = .573, η2p = .003. The main effects of politics were similar in the Control (*F*(1, 545) = 26.38 *p* < .001, η^2_p = .046, 95% C.I. [.018, .085]) and the General Awareness (*F*(1, 617) = 30.33, *p* < .001, η^2_p = .047, 95% C.I. [.020, .083]) condition. The main effect of attractiveness in the Control condition (*F*(1, 545) = 58.91, *p* < .001, η2p = .098, 95% C.I. [.055, .146]) was, if anything, slightly smaller than in the General Awareness condition (*F*(1, 617) = 104.26, p < .001, η2p = .145, 95% C.I. [.097, .195]).

Participants made aware of a general potential for bias did not show reduced bias on either social category, and the impact of general awareness biases favoring attractiveness (*r* = -.04, 95% C.I. [-.10, .02]) or the political ingroup (*r* = -.004, 95% C.I. [-.06, .05]) fell outside the 95% confidence interval for the impact of raising awareness on reducing a targeted bias reported in the meta-analysis of Studies 1a-2b. This evidence is consistent with the interpretation that raising awareness of the 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 that were more physically attractive and those with whom they shared an ingroup identity. Raising awareness of social categories as a potential source of bias and encouraging fairness was effective at reducing the bias only for named social categories. Unmentioned social categories showed no change in magnitude of judgment bias.

The narrow impact of awareness 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 for judgment. Even so, unless the social category was named explicitly, the intervention had no effect. This pattern held in Study 3 when the intervention warned of different types of applicants being unfairly judged, encouraging avoidance of bias toward any irrelevant social categories. The specificity of this effect has important theoretical and practical implications.

Implications

These findings are consistent with models of bias correction suggesting that awareness of potential bias is necessary for reducing judgment biases (Wilson & Brekke, 1994; Wegener & Petty, 1996), 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 made aware of the social category (e.g., political party) rather than relying on a general appeal for avoiding bias. As such, this suggests that effective interventions must invoke the specific social categories to meet the necessary conditions for bias correction (Wilson and Brekke, 1994) or to initiate motivational processes for correcting bias limited to those specific social categories (Moskowitz & Li, 2011). This also suggests that reminding decision-makers of one potential bias may not easily provoke spontaneous introspection and generalization to other sources of bias (Wilson and Brekke, 1994). At least, in the present conditions, the interventions were not sufficient to provoke such generalization. People may need more assistance in identifying the potential sources of their biases.

Simultaneously, decision-makers required relatively little information for awareness manipulations to be effective. A simple warning identifying the social group and statement of avoiding bias toward that group was sufficient to reduce bias (see also Axt, 2017; Carnes et al., 2012; Pope, Price, & Wolfers, 2016). The intervention provided no concrete strategies on how to counteract bias or reminders to avoid bias. The intervention was extremely brief and timed to immediately precede the judgment (cf. Devine et al, 2017). 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.

Gender as a Third Category of Social Bias

Stimuli included men and women in a fully crossed design with attractiveness and ingroup identity. It did not occur to us until after conducting the studies 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 on the JBT in control conditions, such that female applicants received lower criterion than male applicants (Hedge's g = .23, see online supplement for analysis). This gender bias was not much smaller than the biases by attractiveness (Hedge's g = .33) and ingroup status (Hedge's g = .29).

Did awareness interventions about attractiveness, ingroup, or bias in general reduce this gender bias? No. Combining across Studies 1a-2b, awareness of biases in one or more of the other social categories did not impact the gender bias (Hedge's g = -.02). Making participants aware of bias in general in Study 3 also did not alter gender bias (d = -.04). These results are consistent with the confirmatory findings that awareness of the social category is necessary for bias reduction. However, we do not have evidence in these data that drawing attention to gender in the intervention would reduce gender bias. It is conceivable that gender biases are not amenable to awareness interventions. Nevertheless, these exploratory results are consistent with the conclusion that awareness of a specific social category is necessary for bias reduction.

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 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.

Second, it is possible that this conclusion is constrained to features of the intervention. For example, it is possible that other interventions do not require awareness of 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 mindset capable of impacting 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 awareness of the social category to evaluate whether it is necessary for intervention effectiveness, and none have tested whether the interventions generalize to unmentioned social categories.

Third, it is possible that this conclusion is constrained to idiosyncratic features of the selected social categories (attractiveness, institution affiliation, political affiliation, gender) or to interactions between these categories. For example, attitudes toward the attractiveness and ingroup categories were uncorrelated (aggregate *r* = .03), perhaps reducing the likelihood that an intervention targeting one would impact the other. If the biases are correlated, then it is easy to assume that an intervention targeting one might generalize to the others. Of interest would be whether some correlation is sufficient to invoke spontaneous awareness and generalization to make an intervention targeting one social category equally effective at reducing bias toward the other. Likewise, some categories (e.g., race) may be particularly prone to spontaneous awareness whenever potential bias is mentioned, meaning some interventions may be effective for race (and other salient categories) without needing to explicitly identify the social category.

Fourth, it is possible that this 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 have been mitigated. Nevertheless, reliable biases were observed. If anything, we expect that other sampling strategies may reveal larger biases. An open question, however, is whether the intervention strategies will be any more or less effective in other sampling contexts.

Fifth, we did not evaluate whether there is an upper bound of awareness of social categories. Our interventions named one, two or “general” categories of potential bias. Would the intervention be as effective in reducing judgment biases when naming 5, 10, or 20 social categories? It would be surprising if inoculation against judgment bias had no upper bound for named social categories. Yet, we do not have any theoretical basis for identifying what that upper limit is. A productive starting place would be to conduct interventions mirroring standard

EEOC language for protected classes for the practical consideration of the potential effectiveness of those instructions.

Finally, our awareness intervention occurred immediately prior to making judgments. 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 on judgments or evaluations (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 awareness intervention reduced social biases in judgment, but only when alerting participants to the specific social category impacting evaluation, and did not affect other social biases impacting judgment. This implies that awareness of the social category at the time of judgment 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 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.

References

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

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. (2017). The Judgment Bias Task: A reliable, flexible method for assessing individual differences in 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(5), 809-829.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2017). Redefine statistical significance. *Nature Human Behaviour*. DOI: 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*(3), 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*(3), 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*(6), 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.

Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin*, *111*(2), 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*.

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*(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*(2), 212-227.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*(7), 493-503.

Greenman, E., & Xie, Y. (2008). Double jeopardy? The interaction of gender and race on earnings in the United States. *Social Forces*, *86*(3), 1217-1244.

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

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*(2), 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*(11), 1437-1452.

Hosoda, M., & Stone-Romero, E. (2010). The effects of foreign accents on employment-related decisions. *Journal of Managerial Psychology*, *25*(2), 113-132.

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology*, *41*(3), 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., ... & Frazier, R. S. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., ... & Simon, S. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 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*(5), 1029-1050.

Moskowitz, G. B., & Li, P. (2011). Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *Journal of Experimental Social Psychology*, *47*(1), 103-116.

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(1), 167.

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*(2), 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*(2), 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: 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*(6), 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

Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the implicit association test. *Group Processes & Intergroup Relations*, *10*(3), 359-372.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*(3), 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*(5), 506-514.

Schul, Y. (1993). When warning succeeds: The effect of warning on success in ignoring invalid information. *Journal of Experimental Social Psychology*, *29*(1), 42-62.

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology*, *56*(4), 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*(6), 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*(2), 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*(4), 427-439.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117-142.

Woodhams, C., Lupton, B., & Cowling, M. (2015). The snowballing penalty effect: multiple disadvantage and pay. *British Journal of Management*, *26*(1), 63-77.

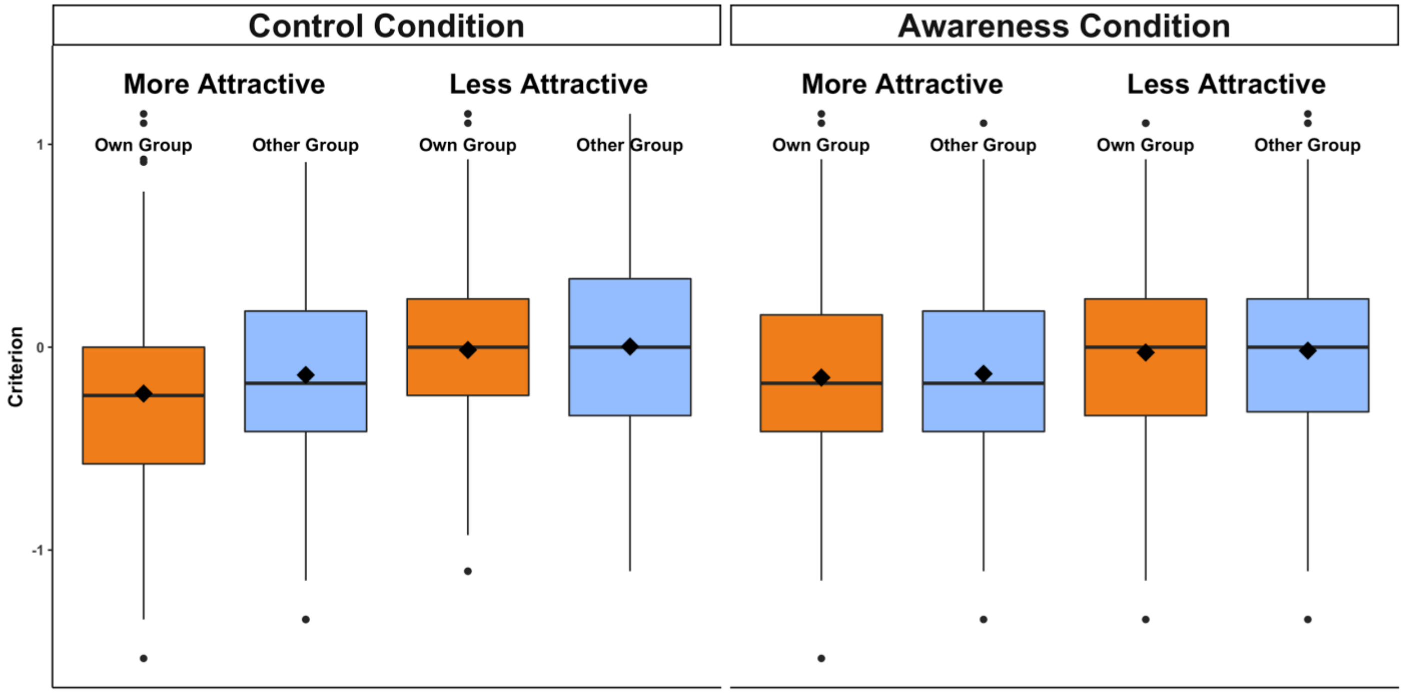

Figure 1. Box plots of criterion values in Study 1a for each experimental condition. Diamond (♦) denotes mean. Own group were applicants from the decision maker's university, and other group were applicants from another university. In the awareness condition, participants were made aware of the possibility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

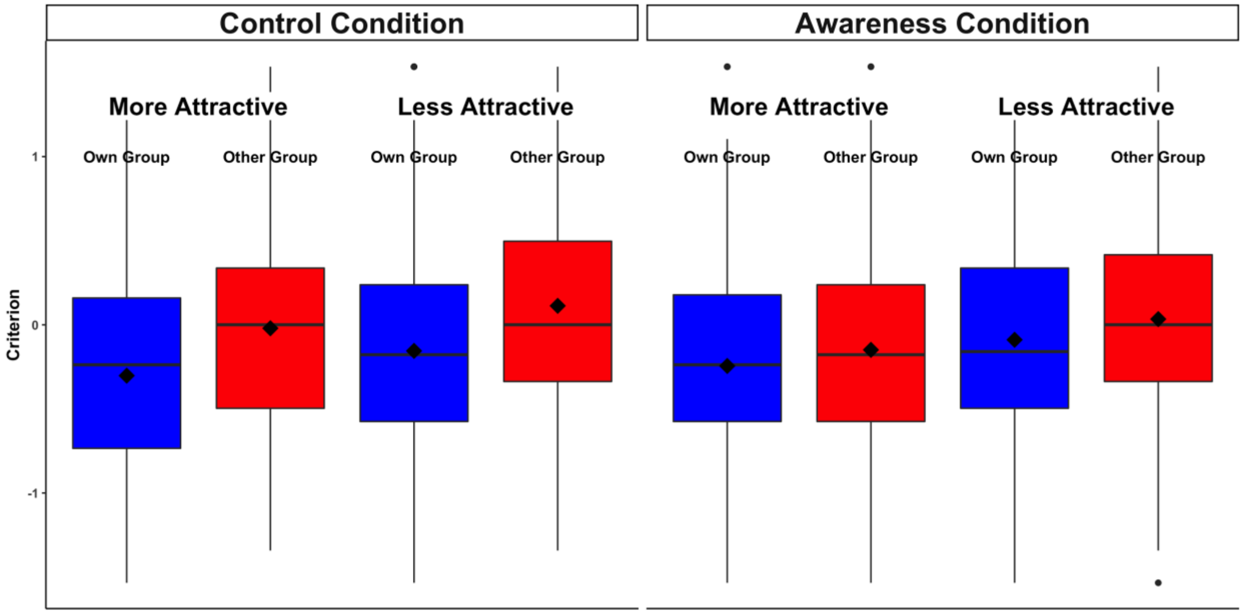

Figure 2. Box plots of criterion values in Study 1b for each experimental condition. Diamond (♦) denotes mean. Own group were applicants from the decision maker's political party, and other group were applicants from the other political party. In the awareness condition, participants were made aware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

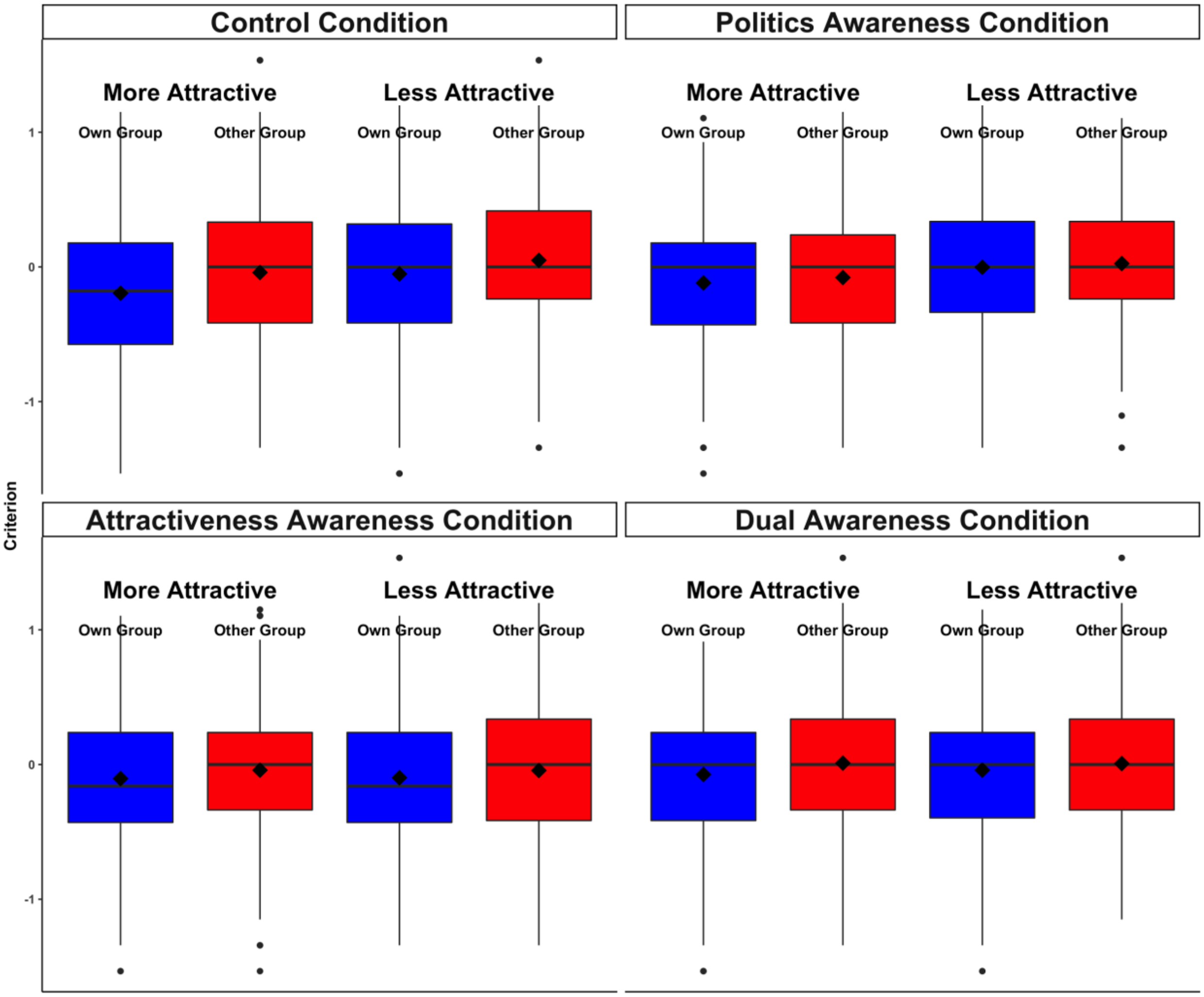

Figure 3. Box plots of criterion values in Study 2a for each experimental condition. Diamond (♦) denotes mean. Own group were applicants from the decision maker's political party, and other group were applicants from the other political party.

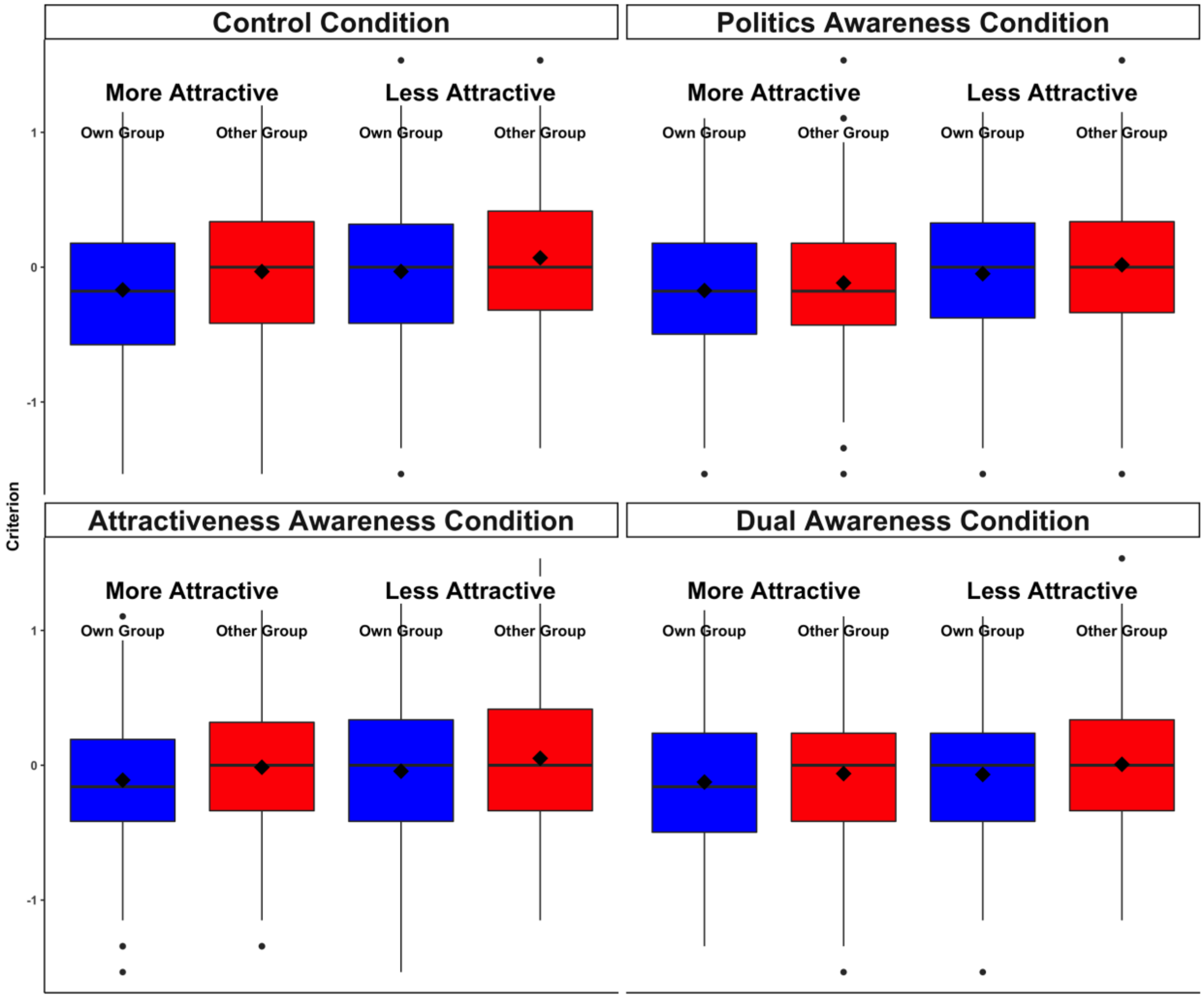

Figure 4. Box plots of criterion values in Study 2b for each experimental condition. Diamond (♦) denotes mean. Own group were applicants from the decision maker's political party, and other group were applicants from the other political party.

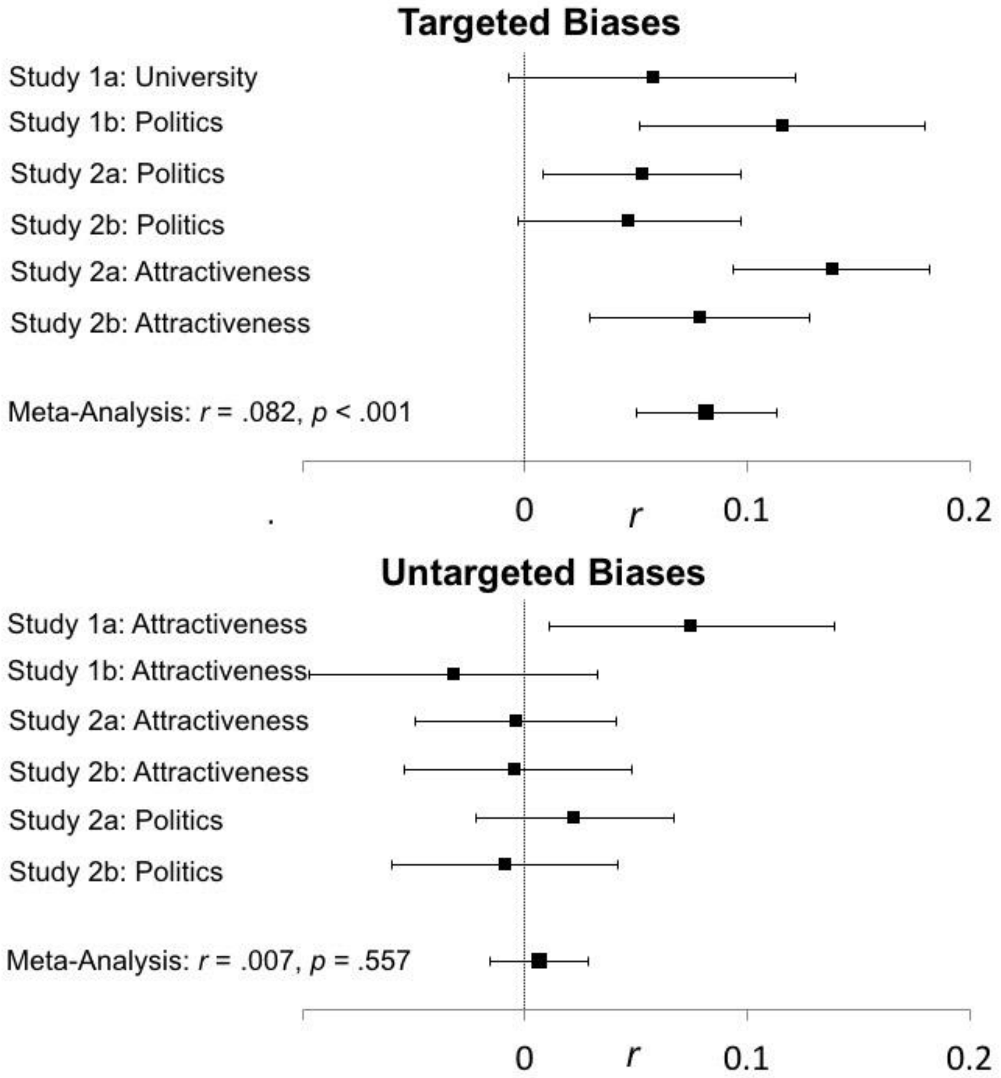

Figure 5. Forest plot and meta-analysis for the impact of awareness on targeted and untargeted biases in Studies 1a-2b. More positive values indicate that the awareness manipulation reduced criterion bias on that category relative to the control condition. Error bars denote 95% confidence intervals on the correlation.

Table 1

Sample Sizes, Criterion Means and Standard Deviation) for All Studies

	More Attractive		Less Attractive	
	Own Group *M* (*SD*)	Other Group *M* (*SD*)	Other Group *M* (*SD*)	Other Group *M* (*SD*)
Study 1a Condition				
Control (*N* = 425)	-.23 (.43)	-.14 (.45)	-.01 (.43)	.003 (.45)
Awareness (*N* = 489)	-.15 (.42)	-.13 (.42)	-.03 (.44)	-.02 (.43)
Study 1b Condition				
Control (*N* = 467)	-.30 (.59)	-.02 (.62)	-.16 (.62)	.11 (.66)
Awareness (*N* = 433)	-.25 (.57)	-.15 (.60)	-.09 (.58)	.04 (.61)
Study 2a Condition				
Control (*N* = 462)	-.20 (.49)	-.04 (.50)	-.05 (.53)	.05 (.51)
Politics Awareness (*N* = 461)	-.12 (.47)	-.08 (.45)	-.002 (.46)	.03 (.45)
Attractiveness Awareness (*N* = 469)	-.10 (.51)	-.04 (.48)	-.10 (.48)	-.04 (.49)
Dual Awareness (*N* = 474)	-.07 (.47)	.01 (47)	-.04 (.49)	.01 (.49)
Study 2b Condition				
Control (*N* = 362)	-.17 (.52)	-.03 (.51)	-.03 (.50)	.07 (.53)
Politics Awareness (*N* = 383)	-.17 (.48)	-.12 (.48)	-.05 (.50)	.02 (.46)
Attractiveness Awareness (*N* = 379)	-.11 (.47)	-.01 (.51)	-.04 (.50)	.05 (.51)
Dual Awareness (*N* = 395)	-.12 (.49)	-.06 (.49)	-.07 (.50)	.01 (.49)
Study 3 Condition				
Control (*N* = 546)	-.16 (.52)	-.04 (.51)	-.02 (.52)	.06 (.51)
General Awareness (*N* = 618)	-.17 (.50)	-.06 (.50)	-.01 (.50)	.08 (.50)