The Impact of Awareness on Reducing Social Bias in Behavior

Jordan Robert Axt

Denver, Colorado

Bachelor of Arts, Duke University, 2008

Master of Social Sciences, University of Chicago, 2010

Master of Arts, University of Virginia, 2014

A Dissertation Presented to the Graduate Faculty of the University of Virginia

in Candidacy for the Degree of Doctor of Philosophy

Department of Psychology

University of Virginia

May, 2017

Abstract

Social bias in behavior is widespread. Interventions to reduce biased behavior have mostly been tested in isolation. As a result, it is unclear whether such interventions derive their effectiveness from theoretically-unique or common mechanisms. In this dissertation, a simultaneous test of four prominent bias reduction interventions revealed that each was able to reduce favoritism toward physically attractive people in an academic admissions task (Study 1). These interventions shared two features: 1) raising awareness that applicants differ on an irrelevant social dimension, and 2) asking participants to behave fairly. When testing these shared features, only increasing awareness reduced biased judgment (Study 2). Moreover, Study 1 interventions were only effective at decreasing socially biased behavior when they raised awareness about the irrelevant social dimension (Study 3). Finally, increased awareness about one irrelevant social dimension had no impact on biased judgment on another irrelevant social dimension (Studies 4a and 4b). Bias reduction strategies with different theoretical origins may actually operate via a shared mechanism of awareness, and effectiveness of awareness interventions may be limited to the reducing bias toward that category rather than engaging general bias reduction decision making strategies.

Word count: 187

Table of Contents

The Impact of Awareness on Reducing Social Bias in Behavior

Jim Everett, a basketball walk-on at North Carolina from 1999 to 2001, actually had a chance to work out his Duke hatred on the basketball court. But even now, as the head of U.S. equity trading for Citigroup, he still can't let it go. While he visits Chapel Hill to recruit Tar Heels for the bank, he says he refuses to visit the school down the road. "There's too many of them up here already," Everett said, referring to the number of Dukies on Citigroup's trading floor." Wall Street Journal, The Complicated Politics of Hating Duke, March 16, 2016.

Social bias in judgment is pervasive. Sometimes, biases align with attitudes and intentions, and are explicitly embraced in decision-making, such as in Mr. Everett's disdain for Duke leading to his unwillingness to hire Duke alumni. In other cases, social biases can influence behavior in unnoticed and undesired ways, leading to actions that contradict attitudes and values. The many documented instances of social biases in academic (e.g., Munro, Lasane & Leary, 2010), political (Lelkes & Westwood, 2017), employment (Pager, 2003), medical (McManus et al., 1995), economic (Doleac & Stein, 2013), and housing (Bartos et al., 2013) contexts suggests that such behavior may deviate from most individuals' egalitarian attitudes or desires for fair treatment (Bertrand, Chugh, & Mullainathan, 2005). The widespread consequences of social biases in behavior, in addition to their capacity for existing outside of awareness or intention (e.g., Hansen et al., 2014), has made them a popular topic of research (e.g., Bertrand & Duflo, 2016; Kaas & Manger, 2012; Moss-Racusin et al., 2012; Milkman, Akinola & Chugh, 2012).

Identifying interventions that can reduce unwanted judgment bias carries both practical and theoretical importance. Prior research has identified numerous interventions that may lessen the impact of unwanted social biases in judgment. However, the existing work is spread across different social domains (e.g., race, gender, age), and often only tests a single treatment condition

against an inactive control condition. As a result, there is support for many ways to reduce biased behavior, but little evidence of comparative effectiveness across interventions and insight into whether these interventions operate via distinct or shared mechanisms.

The present work begins by investigating the effectiveness of multiple interventions to reduce biased social judgment--raising awareness of potential bias in behavior, creating implementation intentions, committing to objectivity, and increasing accountability--with the same sample, procedure, and outcome measure. This approach provides comparative evidence of the effectiveness of these interventions (Lai et al., 2014), and such comparative data provide the basis for hypothesizing about shared or distinct mechanisms for effectiveness. For instance, if some of the interventions are successful at reducing biased behavior but others are not, it would suggest that characteristics present in the effective interventions were critical for eliminating biased behavior in this context. Alternatively, if the interventions are similarly successful at reducing biased behavior, it could either suggest that effectiveness is due to mechanisms specific to the intervention, or mechanisms *shared* across the interventions. Comparative study is an efficient means to investigate mechanisms across diverse intervention strategies.

Study 1 tests four interventions that, based on prior research, could reduce socially based behavior. The selected interventions are not an exhaustive list of bias reduction strategies or even a representative sample of such strategies. Rather, interventions were selected based on prominence in the research literature and adaptability to numerous social domains and judgment contexts. We also selected interventions that explicitly targeted participants' mindsets or strategies, as opposed to more subtle strategies that manipulate the decision context (e.g., placing applicants side by side; Bohnet, van Geen, & Bazerman, 2015) or changing a psychological process that may impact biased judgment (e.g., shifting implicit attitudes; Hsueh, Yogeeswaran,

& Malinen, 2014, Kawakami et al., 2008, Kiefer, Sekaquaptewa, & Barczyk, 2006). The interventions are: (1) raising awareness about the presence of bias; (2) changing choice strategies through implementation intentions; (3) committing beforehand to making objective judgments; (4) increasing feelings of accountability.

Raising awareness. A paper first posted in 2007, and widely covered in the news media, found a racial bias in foul calls among NBA referees in which White players were less likely to receive fouls than Black players (Price & Wolfers, 2010). A later paper looking at the subsequent 2007-2010 seasons found that the racial bias disappeared (Pope, Price, & Wolfers, 2016). Additional analyses suggested that this change in referee behavior could not be explained by referees high in racial bias retiring, the racial makeup of referees changing over time, or new forms of referee training being instituted. Rather, the authors argue that the widespread media attention for the original study created awareness in referees of their racial biases and increased vigilance that then eliminated biased behavior.

Such work suggests that heightened awareness of bias can alter behavior. Indeed, increasing awareness of potential bias in behavior is perhaps the most frequently suggested intervention to combat biases, even ones that may operate outside of conscious awareness or intention (Casey, Warren, Cheesman, & Elek, 2012; Grewal, Ku, Girod, & Valantine, 2013; Handelsman & Sakraney, 2015; Staats, Capatosto, Wright, & Contractor, 2015). An emphasis on raising awareness arises partly from the intuitive appeal of the idea but also from previous basic research suggesting that increased attention to an object or task is associated with better perception (Carrasco, Ling, & Read, 2004), greater effort (Reynolds, Pasternak, & Desimone, 2000), and heightened performance (Spitzer, Desimone, & Moran, 1984). As a result, greater

awareness of potential biases in behavior may create more attention to the sources of these biases, allowing people to better align behavior and intentions.

Theories on debiasing likewise argue that conscious awareness of potential bias is a necessary step to reduce socially biased behavior. For example, in Wilson & Brekke's (1994) model of "mental contamination", people must first gain awareness of the unwanted psychological processes shaping their behavior before they can correct their judgments. Likewise, in the Flexible Correction Model (Wegener & Petty, 1996), individuals must be both motivated to correct bias and aware that a bias is operating in order to lessen biased behavior.

As seen in the study of referee behavior, there is some correlational and experimental work to suggest that raising awareness of potential biases in behavior can reduce bias in behavior. In another study, educating participants about the non-conscious forces that impact behavior reduced biases in self-perception (Pronin & Kugler, 2006). However, experimental evidence on the effectiveness of awareness interventions in intergroup behavior has produced mixed results. Warning participants beforehand that their ratings of an instructor may be susceptible to a "halo effect" (Nisbett & Wilson, 1977), where judgment is influenced by irrelevant information, were just as susceptible to the bias as participants receiving no instructions (Wetzel, Wilson, & Kort, 1981). Similarly, telling participants beforehand about upcoming false feedback did not lessen the impact of such feedback on judgment (Wegner, Coulton, & Wenzlaff, 1985).[1] Yet in other cases, heightened awareness of bias has effectively altered judgments or evaluations in studies of impression formation (Golding, Fowler, Long, & Latta, 1990; Schul, 1993) and attitude change (Petty & Cacioppo, 1986).

[1] Studies arguing in support of the null hypothesis are hard to evaluate considering their relatively small sample sizes (n = 31 per condition in Wetzel et al., 1991 and n = 23 per condition in Wegner et al., 1995). For instance, these samples would provide 28% and 22% power respectively for the effect of awareness found in Study 1 (d = .36).

Implementation intentions. Bias in behavior may be reduced by providing people with a strategy on how to avoid unnecessary social information. Specifically, "implementation intentions" (Gollwitzer, 1999) are concrete plans meant to guide behavior in certain situations, an approach that can then be used to combat the influence of automatic reactions. For example, participants asked to write a take-home essay were more likely to complete the assignment when made to create implementation intentions that outlined when and how they would work on the project (Gollwitzer & Brandstatter, 1997).

Implementation intentions have also been applied to reducing intergroup biases present in implicit associations. Participants provided with an implementation intention where they were asked to repeat to themselves, "I definitely want to respond to the Black face by thinking 'good'" showed less pro-White implicit attitudes on the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) when measured immediately (Lai et al., 2014), though possibly not after 24 hours or more (Lai et al., 2016).[2] Likewise, rehearsing implementation intentions such as "Whenever I see a Black face on the screen, I will think the word, 'safe'" reduced associations between Black and danger on the Weapons Identification Task (Stewart & Payne, 2008).

Similar approaches suggest that implementation intentions can alter both bias-related motivations (Trawalter, Richeson, & Shelton, 2009) and reduce bias in behavior. For example, participants told to adopt the strategy "If I see a person, then I will ignore his race!" made fewer errors on the First-Person Shooter Task (FPST; Correll, Park, Judd, & Wittenbrink, 2002), with process dissociation analyses suggesting that implementation intentions were effective by

[2] As part of a study that compared the effectiveness of multiple interventions on reducing racial bias on the IAT, Lai et al. (2016) found that implementation intentions produced a reliable but small effect ($p = .026$, $\eta^2_p = .01$). However, this finding was interpreted as likely a false positive given that it was the only reliable effect of many tests run in the study.

increasing controlled processing and reducing the impact of automatic stereotypic associations (Mendoza, Gollwitzer, & Amodio, 2010).

Committing to objectivity. Committing to a socially unbiased strategy may decrease reliance on social information in behavior. For instance, in the "should-would discrepancies" paradigm, people acknowledge differences in how one should behave versus how they would behave when interacting with members of other social groups (e.g., Monteith & Voils, 1998; Zuwerink, Monteith, Devine, & Cook, 1996). Activating these discrepancies between how one would (or did) versus how one should act can alter subsequent behavior. In one study, participants low in prejudice towards gay people were led to believe that their earlier selection of a straight versus gay applicant was due to anti-gay bias (Monteith, 1993). These participants then reported greater feelings of guilt and annoyance with themselves compared to participants made to feel that their selection was unrelated to the applicant's sexual orientation. Those made to feel that they had acted in a manner inconsistent with their values then spent a longer time reading an essay on how to reduce anti-gay bias in behavior, suggesting that being exposed to an inconsistency between values and actions created greater effort within participants to change future behavior to be more aligned with their egalitarian attitudes.

Highlighting discrepancies between how one should and would act is related to a more widespread need for consistency between one's values and actions (e.g., Cialdini, 1984). People possess a fundamental motive for consistency in their attitudes and behaviors, and will strive to resolve differences between the two once highlighted (e.g., Festinger, 1957). The need for alignment between values and behaviors can perhaps then be leveraged to reduce bias in social behavior by first asking participants to affirm that others should be treated in an unbiased manner. This initial commitment may shape later behavior to be consistent with one's previously

stated egalitarian values. As a result, stating beforehand that one's behavior should be socially unbiased, and acknowledging that using irrelevant social information in decision-making is unacceptable, may increase motivation to focus on relevant criteria and avoid irrelevant social information during judgment.

Increasing accountability. Making people feel more accountable for their actions can decrease biased judgment (Lerner & Tetlock, 1999). For instance, increased accountability has been shown to reduce reliance on heuristics in financial decision-making (Ashton, 1992; Johnson & Kaplan, 1991) and broaden the amount of relevant information used in evaluation (Siegel-Jacobs & Yates, 1996). Heightened accountability can also lessen bias in social judgment. In one study, participants evaluating a hypothetical job applicant made less biased evaluations after being made to feel more accountable by knowing that they would need to later justify their decisions to an experimenter (Webster, Richter, & Kruglanski, 1996).

In studies using a minimal groups paradigm, increased accountability – again created by warning participants that they would later need to justify their decisions -- reduced ingroup favoritism in allocating hypothetical rewards (Dobbs & Crano, 2001; Hawkins, Sinden, & Nosek, in prep). Likewise, in work concerning how positive mood leads to greater reliance on stereotypes (Bodenhausen, Kramer, & Susser, 1994), participants who anticipated needing to justify their decision showed less bias against Hispanic targets when judging the guilt of a hypothetical student accused of assault.

Given prior research suggesting the effectiveness of these bias reduction interventions, the current studies test the impact of the four strategies on lessening two well-known and widespread biases: favoritism towards more physically attractive people (Studies 1-3) and towards members of one's own ingroup (Studies 4a-4b). These attractiveness and ingroup biases

exist over a range of outcomes (e.g., Beehr & Gilmore, 1982; Cash & Kilcullen, 1985; Hosoda, Stone-Romero, & Coats, 2003; Johnson, Podratz, Dipboye, & Gibbons, 2010) and are robust, with meta-analytic estimates of $d = .61$ for attractiveness (Feingold, 1992) and $d = .36$ for ingroup favoritism (Mullen, Brown & Smith, 1992). Moreover, these biases were selected due to their applicability to various sample populations (i.e., neither is dependent on participant characteristics like race, gender, or age) and given previous evidence that they can exist among people reporting a desire to be unbiased and a perception of having behaved fairly (Axt, Nguyen, & Nosek, 2017).

Study 1

In all studies, we assessed social bias using the Judgment Bias Task (JBT; Axt, et al., 2017). In the JBT, participants evaluate a series of profiles for a specific outcome, such as membership to an academic honor society. Each profile is presented with multiple quantified criteria that are relevant for decision-making as well as other social information that is ostensibly irrelevant, such as a face communicating age, gender, race, and attractiveness. Participants are told to weigh all relevant criteria equally when making their judgment. Across profiles, some are constructed to be systematically better than others. By using multiple criteria, the difference between the more qualified and less qualified profiles is relatively difficult to detect, and the JBT assesses how the ostensibly irrelevant social information impacts evaluation. Through a signal detection analysis, researchers can compare the criterion (c) value for each social group, and bias is evident when criterion is lower for profiles assigned to one social category compared to profiles from another social category.

The JBT is a flexible and reliable measure of bias in behavior (median $\alpha = .70$ for criterion threshold towards each social group in Axt, et al., 2017). For example, in Axt et al.

(2017) participants had a lower criterion for admission to an academic honor society for applicants coming from their own versus a rival school, or their own versus a rival political party, meaning that applicants were more likely to be accepted when coming from one's own school or political party, regardless of qualification. Moreover, across four studies using both online and lab samples, the JBT consistently detected a bias with lower criterion for more versus less physically attractive applicants (average d = .37; Axt et al., 2017). In Study 1, we tested the effectiveness of four interventions at reducing the degree of favoritism toward physically attractive people on the JBT.

Method

Participants

Participants completed the study through the research pool at Project Implicit (implicit.harvard.edu). We sought to collect 655 participants for each of the five experimental conditions who completed at least the JBT (total N = 3275). This sample provided more than 99% power for detecting a small (Cohen's d = .2) within-subjects effect size for each condition, and nearly 100% power for detecting the size of the within-subjects effect (d = .31) found in a previous sample of participants from the same source (Axt, et al., 2017; Study 1b). Between conditions, this sample size provides greater than 80% power at detecting a Cohen's d = .155, which would mean an intervention halved the size of the criterion bias found in the previous sample.

Due to random assignment to conditions and that Project Implicit studies at the time were taken down on fixed days of the week, the final sample size was slightly larger: 3,576 participants, volunteered, consented, and completed at least the JBT. Participants provided demographic information when first signing up for the research pool. Among those who

provided data, 63.3% were female and the mean age was 35.6 (*SD* = 15). By race, 70.3% identified as White, 8.7% African American, 4.0% East Asian, 3.1% South Asian, 4.3% as other or unknown, and 6.6% as biracial. By ethnicity, 8.2% percent reported being Hispanic or Latino. Sample sizes vary across tests due to missing data.

Procedure

The study consisted of five components presented in a fixed order. Participants first received the bias intervention, then completed the academic JBT (Axt, et al., 2017), responded to items measuring perceptions of JBT performance, and finally completed explicit and then implicit attitude measures comparing more and less physically attractive people. See https://osf.io/awd7n/ for the study's pre-registration and https://osf.io/z5wws/ for materials, data analysis syntax, and online supplements for all studies. For all studies, reported analyses are confirmatory tests from the pre-registration unless explicitly noted as exploratory.

Academic JBT. In this version of the JBT, participants received instructions that they would be completing a task where they make accept or reject decisions for an academic honor society. Each applicant had four items of relevant information for participants to weigh equally in selecting applicants: Science GPA (scale of 1-4), Humanities GPA (1-4), letter of recommendation quality (poor, fair, good, excellent), and interview score (1-100).

Qualifications were manipulated to produce two levels of applicant quality. Half of the applications were relatively more qualified and half were less qualified. To determine qualification, each of the four pieces of applicant information were converted to a scale with a maximum score of four. 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 to make the maximum score four for all criteria. For each applicant, the four

scores were summed to determine their level of qualification. Less qualified applicants added to 13 and more qualified applicants added to 14. A full list of the criteria for each profile is listed in Appendix A.

Each application was paired with a face that was pre-rated on attractiveness, with half that were rated as more attractive than the other half (d = 2.64; Axt, et al., 2017). Both more and less attractive faces were half female and half male, and all images were of White people who were smiling.

Participants were instructed to accept approximately half of the applicants to the honor society. Before making their judgments, participants viewed each application/face pairing for one second at a time to get an impression of the range of scores for each dimension. This was intended to provide participants with information about the range and variation in scores so that they would be relatively calibrated for accept/reject decisions on the first judgment trial. There were 64 applications, with eight applications for each combination of gender, attractiveness, and qualification (e.g., eight female, qualified, more attractive applications, eight female qualified, less attractive applications). After viewing all the applicants briefly, participants made accept or reject judgments one-at-a-time for each applicant with no time limit.

Experimental conditions. Before first seeing the applicants, participants were randomly assigned to one of five conditions. For all studies, random assignment was determined by Project Implicit's JSExpPlayer software package (version q0.0.p0.3, 2015) that administered the experimental protocol. In the *Control* condition, participants received no additional instructions. In the four experimental conditions, additional instructions were presented before the encoding phase and after the initial instructions.

In the *Awareness + Fairness* condition, participants read:

The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.

Decision makers are frequently too easy on some applicants, and too tough on others. Prior research suggests that decision makers are easier on more physically attractive candidates and tougher on less physically attractive candidates.

Can you be fair toward all applicants and not be biased by attractiveness? 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 more physically attractive over less physically attractive candidates. When you are done, please type this strategy in the box below.

To maximize the effectiveness of this awareness manipulation, we combined raising awareness of the target of potential bias (less attractive people) with instructions to treat all applicants fairly. Prior research suggests that appeals to fairness and social norms of egalitarianism may increase motivation to be unbiased and decrease prejudiced responses (e.g., Monteith, Deneen & Tooman, 1996; Devine, Monteith, Zuwerink, & Elliot, 1991; Sinclair, Lowery, Hardin & Colangelo, 2005). This combined instruction maximized participants' awareness of the potential for bias and the desired behavior. In Study 2, we examined these two components separately.

After reading the instructions, participants then typed their strategy in a text box at the bottom of the screen. Immediately before the testing phase, participants in the *Awareness* condition were also reminded: *Please remember to be fair and avoid favoring more physically attractive over less physically attractive candidates.*

In the *Implementation Intentions* condition, participants read (adapted from Mendoza, Gollwitzer, & Amodio, 2010):

> *The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.*
>
> *Besides differing in qualifications, the applicants may also differ in physical attractiveness. Be careful not to let attractiveness affect your admission decisions.*
>
> *Research has shown that adopting the following mindset can reduce this bias: If I see a person, then I will ignore his or her physical attractiveness!*
>
> *Please repeat this strategy to yourself quietly. When you are done, please also type this strategy in the box below.*

Participants then typed the strategy in a text box at the bottom of the screen. Immediately before the testing phase, participants in the *Implementation Intentions* condition were also reminded: *Please remember to adopt the following mindset when making your decisions: If I see a person, then I will ignore his or her physical attractiveness!*

Participants in the *Commitment* condition were asked to first report how people should behave before starting the task themselves. Specifically, participants were made to feel more motivated to be objective by reading:

> *The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.*
>
> *All applicants will also be shown with an image of their face. Some applicants may be more physically attractive than other applicants.*

Before you begin, you will answer some questions about how you want to perform on the task, and what strategies are appropriate to use.

Afterwards, participants answered three questions that would likely require them to report valuing unbiased behavior: 1) How do you **want** to perform on the task? (-3= "I want to be extremely easier on less physically attractive applicants and extremely tougher on more physically attractive applicants" to +3= "I want to be extremely easier on more physically attractive applicants and extremely tougher on less physically attractive applicants", and 0= "I want to treat more and less physically attractive applicants equally"); 2) How **should** people perform on the task? (-3="People should be extremely easier on less physically attractive applicants and extremely tougher on more physically attractive applicants" to +3= "People should be extremely easier on more physically attractive applicants and extremely tougher on less physically attractive applicants", and 0 = "People should treat more and less physically attractive applicants equally"); and 3) Do you think it is appropriate to use physical attractiveness when making admissions decisions? ("Yes" or "No").

Immediately before the testing phase, participants in the *Commitment* condition were also reminded: *Please remember your responses concerning how you want to perform on the task, how people should perform on the task, and whether it is appropriate to use physical attractiveness when making admissions decisions.*

Finally, participants in the *Accountability* condition read the following (adapted from Webster, Richter, & Kruglanski, 1996):

The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.

All applicants will also be shown with an image of their face. Some applicants may be more physically attractive than other applicants.

Please accept the most qualified and reject the least qualified applicants. To ensure that you have tried to accept the most qualified applicants, at the end you will write a few sentences explaining your strategy and justifying why that strategy was appropriate. Researchers will later review what you have written, and will examine whether your selections were influenced by physical attractiveness instead of just the qualifications.

Immediately before the testing phase, participants in the *Accountability* condition were also reminded: *Please remember that, after the task, you will explain your strategy. This will be reviewed by the researchers who will also see whether your decisions were influenced by physical attractiveness.* After the testing phase, participants in the *Accountability* condition then wrote a few sentences describing the strategy they used when evaluating applicants.

Within each condition, participants were randomly assigned to one of 16 JBT orders. Across the sixteen orders, each application was equally likely to be paired with a more or less physically attractive face.

Perceptions of JBT performance. Participants answered two questions about task performance. Perceived task performance was measured by the item, "Which statement best describes your performance on the task?" (-3 = "I was extremely easier on less physically attractive applicants and extremely tougher on more physically attractive applicants", +3 = "I was extremely easier on more physically attractive applicants and extremely tougher on less physically attractive applicants", 0 = "I treated more and less physically attractive applicants equally"). Desired task performance was measured by the item, "Which statement best describes how you wanted to perform on the task?" (-3 = "I wanted to be extremely easier on less

physically attractive applicants and extremely tougher on more physically attractive applicants", +3 = "I wanted to be extremely easier on more physically attractive applicants and extremely tougher on less physically attractive applicants", 0 = "I wanted to treat more and less physically attractive applicants equally").

Explicit evaluations of more and less attractive people. Explicit physical attractiveness attitudes were measured by the item, "Which statement best describes you?" (-3 = "I strongly prefer less physically attractive to more physically attractive people", +3 = "I strongly prefer more physically attractive to less physically attractive people", 0 = "I prefer more and less physically attractive people equally").

Implicit evaluations of more and less attractive people. Implicit physical attractiveness attitudes were measured through a four-block, good-focal Brief Implicit Association Test (BIAT; Sriram & Greenwald, 2009; See Appendix B for details of the procedure). The targets were "More attractive people" and "Less attractive people". Stimuli consisted of the two male and two female faces from the decision-making task that were pre-rated as the most and least attractive. The categories were "Good words" (love, pleasant, great, wonderful) and "Other words", which were "Bad words" (hate, unpleasant, awful, terrible). Participants were randomly assigned to complete one of two orders. The BIAT was scored using the *D* algorithm (Nosek et al., 2014).

Results

An error in a software update used across all Project Implicit studies meant that data were not recorded for four trials in the JBT during the last two weeks of data collection ($N = 1206$). For these participants, we calculated JBT outcome variables without these four trials (i.e., 60 total trials). Following Axt et al. (2017), participants were excluded from analysis for accepting less than 20% or more than 80% of the applicants, or for accepting every more attractive or less

physically attractive applicant. 322 participants (9.0%) were excluded based on these criteria. 73 additional participants (2.5% of those completing the BIAT) were excluded from analyses involving the BIAT for having more than 10% of responses faster than 300 ms, as recommended by Nosek et al (2014).

Accuracy was defined as accepting more qualified candidates and rejecting less qualified candidates. Among eligible participants, overall accuracy on the JBT was 67.2% (*SD* = 8.5). The average acceptance rate was close to the recommended 50% (*M* = 51.6%, *SD* = 12.8).

Criterion Bias in Decision-Making

Using signal detection analysis, the JBT assesses participants on their sensitivity (d') and criterion (c) when evaluating profiles. Sensitivity refers to an individual's ability to distinguish between more and less qualified profiles. Higher sensitivity indicates greater ability to accept the more qualified and reject the less qualified profiles. Criterion is the decision threshold, where lower criterion values indicate being more lenient on profiles (i.e., a greater proportion of errors that incorrectly accept less qualified profiles), and higher criterion values indicate being more strict (i.e., a greater proportion of errors that incorrectly reject more qualified profiles). A criterion value of zero indicates that the two errors (accepting less qualified profiles and rejecting more qualified profiles) are equally likely.

We first tested whether there was evidence of differences in criterion between more attractive and less attractive applicants within each condition. Only the *Awareness + Fairness* condition failed to show reliably lower criterion for more relative to less physically attractive applicants, $t(672) = -0.34$, $p = .732$, $d = -0.01$, 95% CI [-0.09,0.06], all other t's > 2.27, p's < .02, d's > 0.09. Conversely, within each condition, only the *Awareness + Fairness* condition had reliable differences in sensitivity between more and less attractive applicants, with slightly

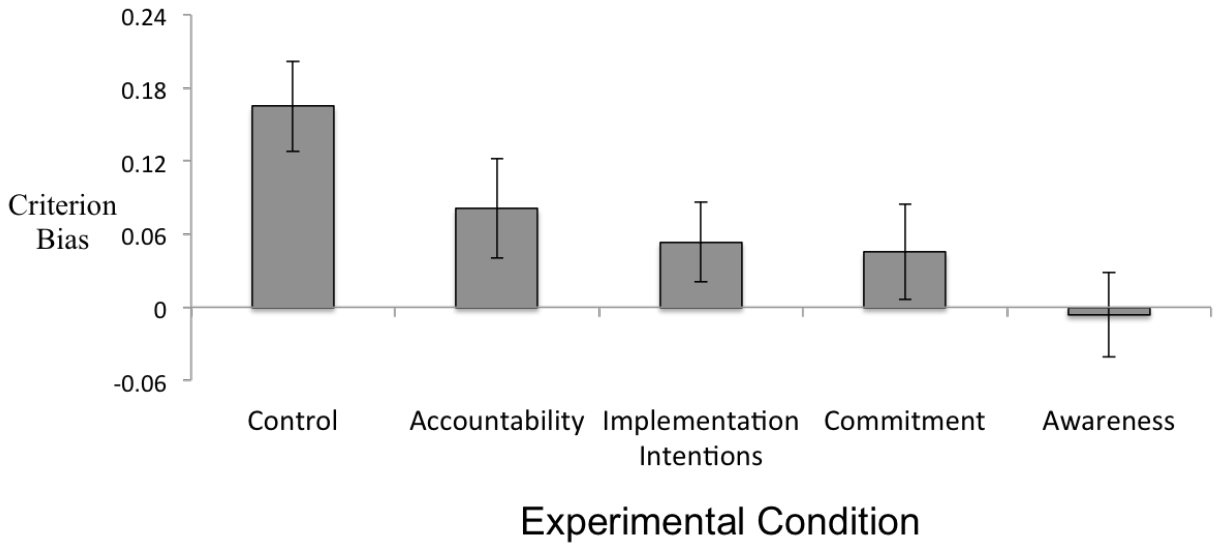

Figure 1. Criterion bias within each condition in Study 1. Higher values mean a lower criterion for more versus less attractive physically attractive applicants. Error bars denote 95% confidence intervals on the mean.

higher sensitivity for less attractive applicants (M = 1.10, SD = 0.65) than more attractive applicants (M = 1.02, SD = 0.63), $t(672) = 2.97$, $p = .003$, $d = 0.11$, 95% CI [0.04, 0.19], all other t's < 1.85, all p's > .07, all d's < 0.07. See Figure 1 for a graphical display of criterion bias (a difference score between the two criterion values such that higher values mean lower criterion for more relative to less physically attractive applicants) in each condition. See Table 1 for means, standard deviations and test statistics for each condition.

We then tested for differences in criterion for more vs. less physically attractive applicants within each condition among participants who reported showing no favoritism on the task, and among those who reported wanting to show no favoritism on the task. Only within the *Control* condition did participants who reported showing no favoritism on the task have lower criterion for more physically attractive relative to less physically attractive applicants, $t(503) = 4.19$, $p < .001$, $d = 0.19$, 95% CI [0.10, 0.27], all other t's < 1.36, p's > .173, d's < 0.06. However, participants who reported *wanting* to show no favoritism on the task had reliably lower

criterion for more physically attractive relative to less physically attractive applicants in the *Control*, *Accountability* and *Implementation Intentions* conditions, t's > 2.06, p's < .041, d's > 0.09, whereas this did not occur in the *Awareness + Fairness* or *Commitment* conditions t's < 1.33, p's > .185, d's < 0.06. See Table 1 for means, standard deviations and test statistics for each condition.

We next tested whether any experimental conditions differed in levels of the criterion bias difference score relative to the control condition among all eligible participants. Relative to the *Control* condition, each of the four experimental conditions had lower criterion biases, all t's > 2.99, all p's < .003, all d's > 0.17.[3] See Table 2 for t-statistics and effect sizes for the comparisons between each experimental condition and the *Control* condition.

Attitudes, Desired and Perceived Performance

Explicit attitudes indicated preference for more physically attractive people (M = 0.69, SD = 0.99, d = 0.70). Relative to the *Control* condition, there were no reliable differences in explicit attitudes across experimental conditions, all t's < 1.58, all p's > .113, all d's < 0.10. Overall, 89.6% of participants reported wanting to treat more and less attractive applicants equally, and 79.7% indicated having done so.

Implicit attitudes indicated more positive associations towards more physically attractive people (M = 0.61, SD = 0.46, d = 1.31). Relative to the *Control* condition (M = 0.62, SD = 0.45), only the *Awareness + Fairness* condition had a lower BIAT D score (M = 0.56, SD = 0.45), $t(1184) = 2.29$, $p = .022$, $d = .13$, 95% CI [.02, .25], all other t's < 0.43, all p's > .669, all d's <

[3] In an exploratory analysis, we also tested whether there was differential dropout between any of the experimental and control conditions. In the Control condition, 73.1% of participants completed the JBT and 64.4% completed the study, and these rates did not differ from those in the Accountability (70.1% completed JBT, 62.1% completed study), Awareness (72.2% completed JBT, 64.3% completed study), Implementation Intentions (73.1% completed JBT, 65.8% completed study) or Commitment (72.7% completed JBT, 64.0% completed study) conditions, all p's > .125.

0.03. See Table 3 for means and standard deviations of implicit and explicit attitudes within each condition.

We next tested whether any experimental conditions differed from the *Control* condition in reporting showing no bias. Relative to the *Control* condition (76.8%), only the *Implementation Intentions* condition (83.3%) had a higher percentage of participants reporting showing no bias, $\chi^2(1, N = 1301) = 8.40$, $p = .004$, all other $\chi 2$'s < 3.04, all p's > .094. Finally, we tested whether any experimental conditions differed from the *Control* condition in the percentage of participants reporting a desire to show no bias. Relative to the *Control* condition, there were no reliable differences in the frequency of reporting a desire to show no bias, all χ^2's < 2.58, all p's > .123. See Table 3 for the proportion of participants in each condition who reported having been fair and reported wanting to be fair on the JBT.

Predicting Criterion Bias

Using the criterion bias difference score, an analysis across all eligible participants in all conditions found that bias was reliably correlated with BIAT *D* scores ($r = .16$, $p < .001$, 95% C.I. [.13, .20]), explicit preferences for more attractive people ($r = .14$, $p < .001$, 95% C.I. [.10, .17]), perceptions of performance ($r = .20$, $p < .001$, 95% C.I. [.16, .23]), and desired performance ($r = .07$, $p < .001$, 95% C.I. [.03, .10]).

A simultaneous linear regression with implicit attitudes, explicit attitudes and condition (coded with *Control* as the reference) predicting criterion bias revealed that implicit (B = .15, $p <$ .001) and explicit attitudes (B = .05, $p < .001$) were reliably and positively related to differences in response criterion, while experimental conditions were all reliably and negatively related (all B's < -.08, all p's < .007). Another simultaneous linear regression that added perceived and desired performance revealed that implicit attitudes (B = .13, p < .001), explicit attitudes (B =

.04, $p < .001$), and perceived performance (B = .13, $p < .001$) were all positive, unique predictors of criterion bias, while experimental conditions were all negative, unique predictors (all B's < -.07, all p's < .009), and desired performance (B = .02, p = .393) was not a unique predictor. These variables accounted for 9.1% of the attractiveness difference in criterion bias. Finally, we conducted a linear regression using all of the above variables and interactions with experimental condition (see Table 4 for coefficients and p values for all terms). These variables accounted for 10.7% of the attractiveness difference in criterion bias.

Discussion

Participants in the *Control* condition replicated past work by showing an automatic and unintended criterion bias favoring more over less physically attractive applicants (Axt, et al., 2017). However, all four of the interventions reduced the criterion bias, and did so with relatively comparable effectiveness (reduction in bias effect sizes: *Accountability* $d = .17$, *Awareness + Fairness* $d = .36$, *Commitment* $d = .24$, *Implementation Intentions* $d = .24$). The *Awareness + Fairness* condition was the only intervention to completely debias behavior. These results are notable considering previous interventions to reduce bias in this paradigm were ineffective -- including placing applicants side by side in pairs (Axt, et al., 2017), or rewarding higher accuracy through donation to a charity of a participant's choice (Axt, Ebersole, & Nosek, 2016). There was some but not consistent evidence that the interventions changed participants' implicit or explicit attitudes toward physically attractive people and their perceived or desired performance on the task. Those effects were inconsistent and weak enough that they require replication before taking seriously.

If only a subset of the interventions had reduced criterion bias on the JBT, it would suggest that details of the procedure specific to the effective interventions were critical for

lessening social bias in behavior. However, all of the Study 1 interventions were effective at reducing criterion bias on the JBT. Although each of the four interventions could have been effective through mechanisms specific to that manipulation, one intriguing possibility is that the four manipulations' effectiveness is a consequence of the *same* mechanism(s), and that aspects shared across the interventions were responsible for eliminating biased behavior. In fact, the *Awareness + Fairness* condition was the theoretically simplest and most effective intervention, and all the interventions included some aspect of increasing awareness. This outcome raises the possibility that the added features of the other experimental conditions were *diluting* rather than strengthening their effectiveness. However, it could also be that components specific to the *Awareness + Fairness* manipulation were responsible for its debiasing effect.

In most research applications of social judgment bias, interventions are tested against an inert control condition. Many of these interventions are rooted in rich theoretical elaboration about why the experimental manipulation was effective, such as invoking awareness, increasing accountability, creating commitment, or using an implementation mindset. The present comparative evidence introduces the possibility that a common theoretical explanation could account for all interventions' effectiveness. Of course, it is also possible that each intervention employed unique mechanisms for effectiveness. Study 2 extends this line of reasoning by manipulating specific features common to all four Study 1 interventions as an initial step toward evaluating the extent to which shared features could account for bias intervention effectiveness.

Study 2

If the effectiveness of the Study 1 interventions was due to components shared across the manipulations rather than features specific to each intervention, then natural follow-up questions are 1) What are these shared components? and 2) Are any of these shared components necessary

or sufficient for intervention effectiveness? In Study 2, I examined the influence of two features that were present in all of the interventions from Study 1: 1) Asking participants to be fair towards all applicants; and 2) Raising awareness that physical attractiveness varies across applicants. This design can then test whether one factor alone or both are necessary to reduce bias in social judgment.

In Study 1, all of the interventions suggested that participants should try to be fair in their evaluation of applicants. This motivation to be fair and avoid using physical attractiveness during the task took subtly different forms across the four manipulations. In the *Awareness + Fairness* condition, participants were simply challenged to "be fair" when evaluating applicants. In the *Implementation Intentions* condition, participants were asked to "not let physical attractiveness affect your admission decisions". In the *Accountability* condition, participants were asked to "accept the most qualified and reject the least qualified applicants". Finally, the *Commitment* condition asked participants to be fair towards all applicants somewhat indirectly. Participants answered items about how people should perform on the task and whether it was appropriate to use physical attractiveness when making admissions decisions, with 94.1% of participants indicating people should treat more and less physically attractive applicants equally and 93% reporting it is inappropriate to use physical attractiveness when making admissions decisions. Given this shared approach across all Study 1 manipulations, Study 2 tests whether instructions to be fair are sufficient on their own to reduce socially biased judgment, or if effectiveness is contingent on also being aware of the target of possible bias (physical attractiveness).

Every intervention in Study 1 also drew participants' attention to the physical attractiveness of the applicants. In the *Implementation Intentions*, *Accountability*, and *Commitment* conditions, participants were alerted to the fact that applicants will vary in their

physical attractiveness. And, while not stated directly, this may have led participants to the spontaneous conclusion that people tend to favor more attractive people over less attractive people. However, in the *Awareness + Fairness* condition, participants were explicitly alerted to a potential bias favoring the physically attractive. It is possible that identifying the social category is sufficient to account for the effectiveness of raising awareness on reducing social bias toward that category. It is also possible that adding awareness of the potential bias in behavior additionally strengthens the effect of raising awareness. Indeed, Study 1 findings are suggestive of this possibility given that the *Awareness + Fairness* condition produced the largest effect (d = .36) and was the only intervention able to completely debias behavior. As a result, Study 2 uses an adaptation of this stronger *Awareness + Fairness* condition to maximize detectability of an effect.

Using a factorial design, Study 2 examined whether calls for fairness, raising awareness, or both of these factors in combination are effective at reducing favoritism in judgment. One or both of these factors may be necessary to reduce social bias on the JBT, as each were present in the effective manipulations of Study 1 but neither were present in previous ineffective manipulations, such as in placing applicants side by side (Axt, et al., 2017) or incentivizing higher accuracy (Axt, et al., 2016).

There are several possible outcomes. First, results could show only a main effect of awareness, meaning that increased awareness is sufficient to reduce biased behavior and instructions to be fair have no effect. Second, results could show only a main effect of fairness, meaning that instructions to be fair are sufficient to reduce biased behavior and awareness has no effect. Third, there could be main effects of both awareness and fairness, meaning that both interventions independently are sufficient to reduce biased behavior. Fourth, there could be an

interaction between awareness and fairness, suggesting that the combination of awareness and fairness is more effective than the independent contributions of each intervention on its own. In fact, if there were *only* an interaction and no main effects, it could suggest that neither intervention is sufficient on its own to reduce biased behavior.

Method

Participants

Participants completed the study through the research pool at Project Implicit. We aimed for 220 participants per experimental condition who completed at least the JBT. That sample size would provide over 98% power for detecting either main effect or the interaction between fairness and awareness for Cohen's $d = .36$, which was the effect for reduction in criterion bias found between the *Control* and *Awareness* conditions in Study 1. This sample size would also provide 84% power for detecting either main effect and 63% power for detecting the interaction between fairness and awareness for a small effect ($d = .20$). See https://osf.io/2q8vv/ for the study's pre-registration.

Due to eligibility restrictions and random assignment to conditions, the final sample size was slightly larger: 1,162 participants, volunteered, consented, and completed at least the JBT. Among those who provided data, 67.8% were female and the mean age was 33.6 ($SD = 14.9$). By race, 70.5% identified as White, 7.9% African American, 4.0% East Asian, 3.2% South Asian, 4.5% as other or unknown, and 6.0% as biracial. By ethnicity, 10.8% percent reported being Hispanic or Latino. Sample sizes vary across tests due to missing data.

Procedure

The study consisted of five components completed in the same order as Study 1: participants first received the bias intervention, then completed the academic JBT, followed by

items measuring perceptions of JBT performance, and then measures of explicit and implicit attitudes toward physically attractive people.

Experimental conditions. Before completing the JBT, participants were assigned to one of four experimental conditions in a 2 (*Awareness* vs. *No Awareness*) by 2 (*Fairness* vs. *No Fairness*) between-subjects design. In the *No Awareness / No Fairness* condition, participants read the following before the encoding phase and after the initial instructions:

> *The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.*

The remaining three conditions had the same text as above and then text that either raised awareness of variation among targets in physical attractiveness and the potential for bias or asked participants to be fair towards all applicants, or both. In the *Awareness / No Fairness* condition, the following text was added:

> *In addition to differing on their qualifications, candidates will differ in physical attractiveness. Prior research suggests that decision makers are easier on more physically attractive candidates and tougher on less physically attractive candidates.*

In the *No Awareness / Fairness* condition, the following text was added:

> *Can you be fair toward all applicants and use only their academic qualifications? When you make your "Accept" and "Reject" decisions, be as fair as possible. Please tell yourself quietly that you will be fair***.**

Finally, in the *Awareness / Fairness* condition, instructions included both the awareness and fairness text above.

Participants received reminders consistent with their condition immediately before the testing phase. The *Awareness* reminder text read: *Remember that some candidates will be more or less physically attractive than others and that people are easier on more attractive candidates and tougher on less attractive candidates.* The *Fairness* reminder text read: *Please be fair when evaluating the applicants.*

Academic decision-making task. Participants completed the same JBT as in Study 1 with two changes. First, the encoding phase was removed after a separate study found that removing the encoding phase did not reliably influence overall sensitivity or the degree of criterion bias (see Axt, et al., 2017). Second, to provide more information to participants, each of the four academic qualifications were presented with their ranges and median values in the initial JBT instructions.

Perceptions of performance and attractiveness attitudes. Participants completed the same measures of perceived performance, desired performance as well as explicit and implicit attitudes toward physically attractive people as in Study 1.

Results

Participants were excluded from analysis for accepting less than 20% or more than 80% of the applicants, or for accepting every more or less physically attractive applicant. 222 participants (19.1%) were excluded based on these criteria.[4] 32 additional participants (3.9% of those completing the BIAT) were excluded from analyses involving the BIAT for having more than 10% of responses faster than 300 ms.

[4] The higher exclusion rate here is likely due to removing the encoding phase. Whereas inattentive participants were more likely to drop out during the passive encoding phase and not complete the JBT in previous studies, these participants were more likely to finish the JBT when there was no encoding but still exhibited careless responding, such as choosing the same response option for all trials. In text analyses use our pre-registered exclusion criteria, but primary conclusions are not altered when including all participants (analyses available in the online supplement).

Among eligible participants, overall accuracy on the task was 65.6% (*SD* = 9.5). The average acceptance rate was close to the recommended 50% (*M* = 52.5%, *SD* = 13.5).

Criterion Bias in Decision-Making

The primary analysis was a 2 (Awareness vs. No Awareness) by 2 (Fairness vs. No Fairness) ANOVA on the criterion bias difference score. This analysis revealed a reliable main effect of awareness, $F(1,936) = 7.91$, $p = .005$, $\eta^2_p = .008$, 95% C.I. [.001, .024], such that participants who received an awareness manipulation had lower criterion biases (*M* = .09, *SD* = .52) than participants who did not (*M* = .18, *SD* = .51). There was not a reliable main effect of fairness, $F(1,936) = .44$, $p = .509$, $\eta^2_p < .001$, or an interaction between awareness and fairness $F(1,936) = .88$, $p = .347$, $\eta^2_p = .001$ (see Figure 2).

Follow-up analyses showed that participants in all four experimental conditions had reliably lower criterion for more vs. less physically attractive applicants (all *t*'s > 2.36, all *p*'s <.020, all *d*'s > .15), and failed to replicate the Study 1 result that awareness completely debiased behavior. However, among participants who reported wanting to treat and having treated more and less physically attractive applicants equally, only those receiving the awareness manipulation failed to show a reliable bias in criterion. See Table 5 for means, test statistics, effect sizes and confidence intervals.

Attitudes, Desired and Perceived Performance

Explicit attitudes indicated preference for more physically attractive people (*M* = 0.72, *SD* = 0.96, *d* = 0.74) and implicit attitudes indicated more positive associations towards more attractive than less attractive people (*M* = 0.70, *SD* = 0.51, *d* = 1.38). Overall, 87.9% of participants reported wanting to treat more and less attractive applicants equally, and 73.4% indicated having done so.

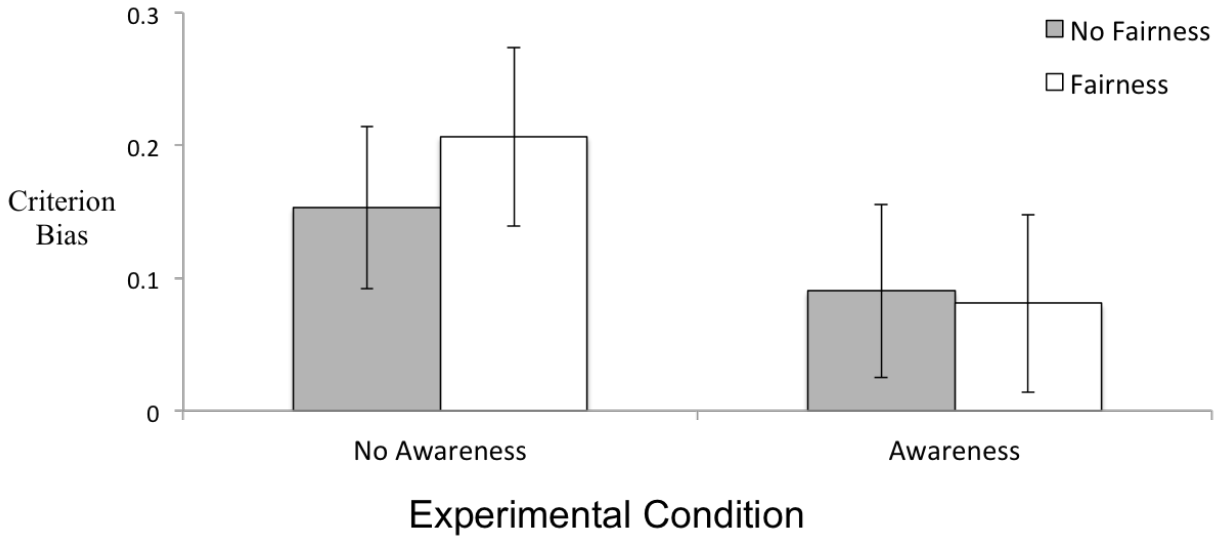

Figure 2. Criterion bias within each condition in Study 2. Higher values mean a lower criterion for more versus less attractive physically attractive applicants. Error bars denote 95% confidence intervals on the mean.

Three 2 (Awareness vs. No Awareness) by 2 (Fairness vs. No Fairness) ANOVAs on perceived performance, desired performance, and explicit attitudes towards more vs. less physically attractive people showed no reliable main effects or interactions between awareness and fairness (all *p*'s >.185). For BIAT *D* scores, there was a main effect of fairness, $F(1,777) = 6.92$, $p = .009$, $\eta^2_p = .009$, 95% C.I. [.001, .026], such that participants who received the fairness manipulation had lower BIAT *D* scores ($M = .66$, $SD = .50$) than participants who did not ($M = .75$, $SD = .52$). There was not a reliable main effect of awareness, $F(1,777) = .01$, $p = .923$, $\eta^2_p < .001$, or an interaction between awareness and fairness $F(1,777) = 3.55$, $p = .060$, $\eta^2_p = .005$. See Table 6 for attitude and performance measure means in each condition and the online supplement for output for each analysis.

JBT Performance Among Participants Not Wanting To Be Fair

Although most participants reported wanting to treat more and less physically attractive applicants equally, we analyzed the impact of the awareness and fairness manipulations among

those who reported not wanting to treat more and less physically attractive applicants equally (N = 105). There was no reliable main effect of either awareness, $F(1,101) = 2.06$, $p = .155$, η^2_p =.020, or fairness, $F(1,101) = .70$, $p = .404$, η^2_p =.007, and no interaction between awareness and fairness, $F(1,101) = 2.32$, $p = .131$, η^2_p =.022. Each of these effect sizes fell within the 95% confidence interval for analyses in the full sample. Given the small sample size and low power for these tests, it is difficult to conclude whether awareness and fairness impacted criterion bias among participants not wanting to be fair, a result that could inform models of biased judgment (e.g., Wegener & Petty, 1996, Fazio, 1990). In the General Discussion, we revisit this question and report a meta-analysis of the impact of awareness among participants not wanting to be fair.

Non-Focal Tests: Predicting Bias in Criterion

The above analyses tested hypotheses central to the study's hypothesis. However, these data also improve estimate precision for questions that other researchers may find interesting but are not central to our study design. For that reason, we only report estimates and confidence intervals (not p-values) for the tests below in an effort to reduce the Type I error rate among the study's focal tests.

Using the criterion difference score, an analysis across all eligible participants in all conditions found that criterion bias was positively but weakly correlated with BIAT D scores (r = .12, 95% C.I. [.05, .18]), explicit preferences for more attractive people (r = .13, 95% C.I. [.06, .20]), and perceptions of performance (r = .27, 95% C.I. [.21, .33]), and was negatively and weakly correlated desired performance (r = -.02, 95% C.I. [-.09, .05]).

A simultaneous linear regression with implicit attitudes, explicit attitudes, awareness condition (coded with *Control* as the reference) and fairness condition (coded with *No Fairness* as the reference) revealed that criterion bias was positively related with implicit attitudes (B =

.11, 95% C.I. [.04, .18]), explicit attitudes (B = .06, 95% C.I. [.02, .10]), and fairness condition (B = .04, 95% C.I. [-.03, .11]), while awareness condition was negatively related (B = -.09, 95% C.I. [-.16, -.02]). Another simultaneous linear regression that added perceived and desired performance revealed that implicit attitudes (B = .11, 95% C.I. [.04, .18]), explicit attitudes (B = .05, 95% C.I. [.01, .08]), perceived performance (B = .18, 95% C.I. [.14, .23]) and fairness condition (B = .03, 95% C.I. [-.04, .10]) were positively related with criterion bias, while desired performance (B = -.05, 95% C.I. [-.12, .02]) and awareness condition (B = -.09, 95% C.I. [-.16, -.02]) were negatively related. These variables accounted for 12.0% of the attractiveness difference in criterion bias. A final linear regression using all of the above variables and interactions with experimental condition accounted for 13.2% of the attractiveness difference in criterion bias (see Table 7 for coefficients and confidence intervals).

Discussion

Replicating Study 1, participants made more aware of differences in targets' physical attractiveness showed lower levels of criterion bias. Merely telling participants to be fair did not reduce biased judgment, nor did adding instructions to be fair to the awareness text increase the intervention's effectiveness. Unlike Study 1, awareness effectively reduced criterion bias but did not completely debias behavior. This finding, combined with the smaller effect size for the main effect of awareness in Study 2 (d = .19) suggest that the impact of the awareness intervention in Study 1 may have been an overestimation of the true effect size. However, the awareness intervention effectively debiased behavior among participants who reported wanting to be fair and a perception that they were fair.

An intervention that only instructed participants to be fair had no discernible impact on JBT performance. The ineffectiveness of the fairness manipulation without awareness of the

target dimension in Study 2 aligns with previous findings that a large majority of participants (for example, 90.3% in Study 1's *Control* condition) already report a desire to be fair towards more and less physically attractive applicants, meaning a motivation for fair treatment is already pervasive. Conversely, the awareness instructions may have been effective in Study 2 because they alerted participants to a social dimension (physical attractiveness) that they either would not notice or would notice but did not believe would impact their behavior.

These data further suggest that all four interventions in Study 1 are effective because they raise awareness about the target of potential bias, and not because of the other theoretically rich features of their methodologies. Study 3 is a more direct test of whether manipulating awareness is sufficient to account for the effectiveness of each of the Study 1 interventions. More specifically, in Study 3, I revisited the interventions of Study 1 and directly manipulated awareness that targets will differ in physical attractiveness in each of them. This produced a 2 (Awareness vs. No Awareness) by 4 (Intervention: Awareness, Accountability, Commitment to Objectivity, Implementation Intention) design.

There are several possible study outcomes. First, if intervention effectiveness is not contingent on raising awareness, then we would observe that all of the interventions would show reduced bias compared to the *No Awareness* condition of the *Awareness* intervention (i.e., the de facto *Control* condition). This would produce an interaction between the intervention and awareness factors. Second, if intervention effectiveness relies on raising awareness, then all of the intervention conditions would be effective with awareness, but not without awareness. This would produce a main effect of awareness. Third, if some interventions rely on awareness of the target dimension and others do not, then we would observe a more complex pattern with some interventions showing reductions in the *No Awareness* condition and others not. This would

produce an interaction between interventions and awareness manipulations. In this case, Study 3 would rule out the possibility that awareness is the sole mechanism supporting effectiveness of these different interventions.

Study 3

Method

Participants

Participants completed the study through the research pool at Project Implicit. We originally aimed to recruit an average of 250 eligible participants per experimental condition, for a total of 2000 participants. This sample would provide over 99% power for detecting main effects and 94% power for detecting an interaction between awareness and intervention type for Cohen's d = .29, which was the weighted average effect for reduction in criterion bias found between the *Control* and *Awareness* conditions in Study 1 and Study 2.

However, this sample size provided relatively low power for specific tests of the effect of awareness within each intervention (e.g., 250 per cell provides 61% power at detecting a small effect of d =.20). As a result, our original pre-registration (https://osf.io/aw9nw/) stated that we would analyze the data after 2000 participants to test if there was a main effect of awareness in the 2 x 4 ANOVA concerning average levels of criterion bias in each condition. If there was no main effect, we would end data collection. If there was a reliable main effect of awareness, we would collect additional data. Results after 2000 participants showed a main effect of awareness, so we pre-registered a new stopping rule (https://osf.io/89ejn/) stating that we would collect data until each within-manipulation contrast had at least 80% power at detecting a small effect (d = .20). To account for the planned data-peeking, I report p-augmented (Sagarin, Ambler, & Lee, 2014), which estimates the range of the possible inflated Type I Error rate (i.e., the probability of

a false positive) that follows collecting and analyzing data multiple times. In most cases where there is only one additional round of data collection, this increase is small and the probability of a false positive remains close to 5%.

In total, 4,116 participants, volunteered, consented, and completed at least the JBT. Among those who provided data, 61.5% were female and the mean age was 36.2 (*SD* = 15.5). By race, 71.7% identified as White, 8.0% African American, 3.4% East Asian, 3.5% South Asian, 4.2% as other or unknown, and 6.0% as biracial. By ethnicity, 8.2% percent reported being Hispanic or Latino. Sample sizes vary across tests due to missing data.

Procedure

The study had a 2 (Awareness vs. No Awareness) by 4 (Intervention: Awareness, Accountability, Implementation Intentions, Commitment to Objectivity) between-subjects design. The study consisted of five components, completed in the following order.

Experimental conditions. Participants were assigned to versions of the *Awareness, Accountability*, *Implementation Intentions* and *Commitment to Objectivity* conditions used in Study 1 that either did or did not raise awareness that applicants differ on physical attractiveness.

Participants in the *Awareness* condition read the following, with the bold and bracketed text indicating the difference between the *Awareness* and *No Awareness* versions:

> *The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.*
>
> *In addition to differing on their qualifications, candidates will differ in* ***[physical attractiveness / other ways].*** *Prior research suggests that decision makers are easier on*

> ***[more physically attractive candidates and tougher on less physically attractive candidates / some types of candidates and tougher on other types of candidates]***.

Before beginning the testing phase, participants also read:

> *Remember to avoid favoring* ***[more physically attractive over less physically attractive candidates / some types of candidates over other types of candidates]***.

In the *Implementation Intentions* condition, participants read:

> *The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.*
>
> *In addition to differing on their qualifications, candidates will differ in [**physical attractiveness / other ways]**. Prior research suggests that decision makers are easier on* ***[more physically attractive candidates and tougher on less physically attractive candidates / some types of candidates and tougher on other types of candidates]***.
>
> *Be careful not to let* ***[attractiveness/ irrelevant information]*** *affect your admission decisions. Research has shown that adopting the following mindset can reduce this bias:*
>
> ***[If I see a person, then I will ignore physical attractiveness!/ If I see a person, then I will ignore irrelevant information!]***
>
> *Please repeat this strategy to yourself quietly. When you are done, please also type this strategy in the box below.*

Participants were then provided with a text box to enter in the strategy they were told to adopt.

Before beginning the testing phase, participants also read:

*Remember to adopt the following mindset when making your decisions:**[If I see a person, then I will ignore physical attractiveness!/ If I see a person, then I will ignore irrelevant information!]***

In the *Commitment to Objectivity* condition, participants read:

The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.

*In addition to differing on their qualifications, candidates will differ in **[physical attractiveness / other ways].** Prior research suggests that decision makers are easier on **[more physically attractive candidates and tougher on less physically attractive candidates / some types of candidates and tougher on other types of candidates]**.*

Before you begin, you will answer some questions about how you want to perform on the task, and what strategies are appropriate to use.

Then, participants answered the following questions:

*1) How do you want to perform on the task? (1 = I do not want to use **[physical attractiveness / irrelevant information]** at all when making admissions decisions.; 4 = I want to use **[physical attractiveness / irrelevant information]** a great deal when making admissions decisions.)*

*2) How should people perform on the task? (1 = People should not use **[physical attractiveness / irrelevant information]** at all when making admissions decisions.; 4 = People should use **[physical attractiveness / irrelevant information]** a great deal when making admissions decisions.)*

3) Do you think it is appropriate to use ***[physical attractiveness / irrelevant information]*** *when making admissions decisions? (1 = Yes, 2 = No).*

Before beginning the testing phase, participants also read:

Remember your responses concerning how you want to perform on the task, how people should perform on the task, and whether it is appropriate to use ***[physical attractiveness / irrelevant information]*** *when making admissions decisions.*

Finally, participants in the *Accountability* condition read:

The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.

In addition to differing on their qualifications, candidates will differ in ***[physical attractiveness / other ways].*** *Prior research suggests that decision makers are easier on* ***[more physically attractive candidates and tougher on less physically attractive candidates / some types of candidates and tougher on other types of candidates]****.*

Please accept the most qualified and reject the least qualified applicants. To ensure that you have tried to accept the most qualified applicants, at the end you will write a few sentences explaining your strategy and justifying why that strategy was appropriate. Researchers will later review what you have written, and will examine whether your selections were influenced by ***[physical attractiveness instead of just the academic qualifications / irrelevant information instead of just the academic qualifications].***

Before the testing phase, participants also read:

Remember that, after the task, you will explain your strategy. This will be reviewed by the researchers who will also see whether your decisions were influenced by ***[physical attractiveness / information other than applicants' academic qualifications]****.*

As in Study 1, participants in the *Accountability* condition wrote a few sentences describing the strategy they used when evaluating applicants following the JBT.

Academic decision-making task. Participants completed the same physical attractiveness JBT used in Study 2.

Perceptions of JBT performance, explicit and implicit attitudes. Participants completed the same measures of perceived performance, desired performance and attitudes as Studies 1 and 2.

Results

Participants were excluded from analysis for accepting less than 20% or more than 80% of the applicants, or for accepting every more or less physically attractive applicant. 644 participants (15.6%) were excluded based on these criteria.[5] 85 additional participants (2.7% of those completing the BIAT) were excluded from analyses involving the BIAT for having more than 10% of responses faster than 300 ms, as recommended by Nosek et al (2014).

Among eligible participants, overall accuracy on the task was 67.2% (SD = 8.8). The average acceptance rate was close to the recommended 50% (M = 52.0%, SD = 12.8).

Criterion Bias in Decision-Making

Our primary analysis was a 2 (Awareness) by 4 (Intervention) ANOVA on the criterion bias difference score. We initially analyzed the data after approximately 2,000 eligible participants to see if there was a main effect of awareness. This first sample (N = 2019) showed a

[5] As in Study 2, the higher exclusion rate in Study 3 is likely due to removing the encoding phase. In text analyses use our pre-registered exclusion criteria, but primary conclusions are not altered when including all participants (analyses available in the online supplement).

reliable main effect of awareness, $F(1, 2011) = 23.97$, $p < .001$, $\eta^2_p = .012$, 95% C.I. [.004, .023], such that participants in *Awareness* conditions had lower criterion biases ($M = .05$, $SD = .48$) than participants in *No Awareness* conditions ($M = .15$, $SD = .45$). There was not a reliable main effect of manipulation, $F(3, 2011) = 0.95$, $p = .416$, $\eta^2_p = .001$, or a reliable manipulation by awareness interaction, $F(3, 2011) = 0.79$, $p = .502$, $\eta^2_p = .001$.

Given the reliable main effect of awareness, we collected additional data until there were enough eligible participants to provide at least 80% power for detecting a small effect ($d = .20$) for the contrast comparing the *Awareness* and *No Awareness* versions within each of the four interventions. Again, to account for analyzing the data multiple times, analyses comparing differences in criterion bias between conditions report *p*-augmented (Sagarin, et al., 2014).

Across all eligible participants ($N = 3469$), a 2 (Awareness) by 4 (Intervention) ANOVA on the criterion bias difference score showed a reliable main effect of awareness, $F(1, 3461) = 36.42$, $p < .001$, $\eta^2_p = .010$, 95% C.I. [.005, .018][6], such that participants in *Awareness* conditions had lower criterion biases ($M = .05$, $SD = .47$) than participants in *No Awareness* conditions ($M = .14$, $SD = .45$). There was not a reliable main effect of manipulation, $F(3, 3461) = 1.43$, $p = .232$, $\eta^2_p = .001$, and no reliable manipulation by awareness interaction, $F(3, 3461) = 1.01$, $p = .388$, $\eta^2_p = .001$. See Figure 3 for a graphical display of criterion bias in each condition.

We then tested for differences in criterion bias between the *No Awareness* and *Awareness* conditions within each intervention. Replicating Study 1, the *Awareness* condition ($M = .03$, $SD = .45$) had a lower criterion bias than the *No Awareness (Control)* condition ($M = .17$, $SD = .43$), $t(901) = 4.55$, $p < .001$, $d = .30$, 95% CI [.17, .43] $p_{augmented} = [0.050, .0500003]$. In the *Accountability* condition, the *Awareness* version ($M = .06$, $SD = .47$) had a lower criterion bias

[6] Given the large sample size and that the original analysis was reliable at $p < .05$, the $p_{augmented}$ criterion for the main effect of awareness was very close to .05, [.050, .0500000000000206].

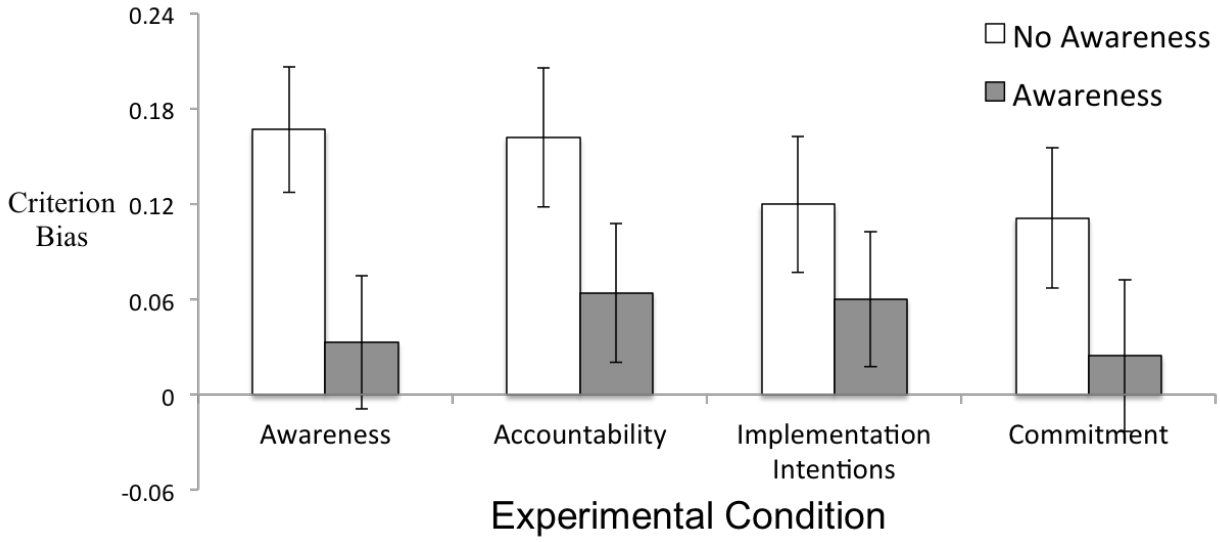

Figure 3. Criterion bias within each condition in Study 3. Higher values mean a lower criterion for more versus less attractive physically attractive applicants. Error bars denote 95% confidence intervals on the mean.

than the *No Awareness* version (M = .16, SD =.46), $t(866) = 3.10$, p = .002, d = .21, 95% CI [.08, .34] $p_{augmented}$ = [.050, .0503]. For the *Objectivity* condition, the *Awareness* version (M = .02, SD =.50) had a lower criterion bias than the *No Awareness* version (M = .11, SD =.44), $t(806) = 2.58$, p = .010, d = .18, 95% CI [.04, .32] $p_{augmented}$ = [.050, .054]. Finally, in the *Implementation* condition, the *Awareness* version (M = .06, SD =.46) had a lower criterion bias than the *No Awareness* version (M = .12, SD =.46), though this contrast was not significant $t(888) = 1.93$, p = .054, d = .13, 95% CI [-.002, .26] $p_{augmented}$ = [.058, .087].

Finally, we tested whether there was evidence of differences in criterion between more attractive and less physically attractive applicants within each condition. Among eligible participants, each condition showed reliably lower criterion for more than less physically attractive participants except for the *Awareness* condition and the *Awareness / Commitment to Objectivity* condition. In a replication of Study 2, among participants reporting a desire to treat more and less physically attractive participants equally (89.0%), all *No Awareness* conditions

showed reliable criterion biases on average whereas no *Awareness* conditions showed reliable criterion biases. This same pattern emerged when looking only at participants reporting a perception of having been fair (77.3%), which also replicated Study 2. See Table 8 for sample sizes, criterion values, *p* values, test statistics and effect sizes.

Attitudes, Desired and Perceived Performance

Explicit attitudes indicated preference for more physically attractive people ($M = 0.72$, $SD = 0.98$, $d = 0.74$) and implicit attitudes indicated more positive associations towards more attractive people ($M = 0.70$, $SD = 0.51$, $d = 1.38$). Overall, 89.0% of participants reported wanting to treat more and less attractive applicants equally, and 77.3% indicated having done so.

Three 2 (Awareness) by 4 (Intervention) ANOVAs on implicit attitudes, explicit attitudes, and desired JBT performance showed no main effects of awareness or manipulation or interactions between awareness and intervention (all F's < 2.02, all p's < .155 η^2_p <.001; full analyses available in the online supplement). However, an ANOVA for perceived performance revealed a small but reliable main effect of awareness, $F(1, 3245) = 5.25$, $p = .022$, $\eta^2_p = .002$, 95% C.I. [.00001, .006] such that participants in *Awareness* conditions perceived their performance to be more fair ($M = .01$, $SD = .71$) than participants in *No Awareness* conditions ($M = .07$, $SD = .74$). There was not an effect of intervention, $F(3,3245) = 2.55$, $p = .054$, $\eta^2_p = .002$, or an intervention by awareness interaction, $F(3, 3245) = 1.81$, $p = .143$, $\eta^2_p = .002$. See Table 9 for means and standard deviations in each condition for perceived performance, desired performance, explicit and implicit attitudes.

Non-Focal Tests: Predicting Bias in Criterion

As in Study 2**,** we pre-registered analyses that were not central to the study's hypothesis but could be of interest for other purposes. For analyses predicting criterion bias, we again only

report estimates and confidence intervals (not *p*-values) in an effort to reduce the Type I error rate among the study's focal tests.

Using the criterion difference score, an analysis across all eligible participants found that criterion bias was positively but weakly correlated with BIAT *D* scores (r = .10, 95% C.I. [.06, .13]), explicit preferences for more attractive people (r = .15, 95% C.I. [.12, .18]), perceptions of performance (r = .24, 95% C.I. [.21, .28]), and desired performance (r = .11, 95% C.I. [.07, .14]).

A simultaneous linear regression with implicit attitudes, explicit attitudes, awareness condition (coded with *No Awareness* as the reference) and intervention condition (coded with *Control* as the reference) revealed that criterion bias was positively related with implicit attitudes (B = .07, 95% C.I. [.03, .10]) and explicit attitudes (B = .07, 95% C.I. [.05, .09]), while awareness condition (B = -.09, 95% C.I. [-.12, -.06]), and the intervention variables (Accountability: B = -.01, 95% C.I. [-.05, .04]; Implementation: B = -.02, 95% C.I. [-.07, .02], Objectivity: B = -.04, 95% C.I. [-.08, .01]) were all small but negatively related.

Another simultaneous linear regression that added perceived and desired performance revealed that implicit attitudes (B = .05, 95% C.I. [.02, .09]), explicit attitudes (B = .05, 95% C.I. [.03, .06]), perceived performance (B = .15, 95% C.I. [.12, .17]), and desired performance (B = .04, 95% C.I. [.002, .07]) were small and positively related to greater criterion biases, while awareness condition (B = -.08, 95% C.I. [-.11, -.05]), and the intervention variables (Accountability: B = -.001, 95% C.I. [-.05, .04]; Implementation: B = -.01, 95% C.I. [-.05, .04], Objectivity: B = -.03, 95% C.I. [-.07, .02]) were all negatively related. These variables accounted for 9.4% of the attractiveness difference in criterion bias. A final linear regression using all of the above variables and interactions with experimental condition (see Table 10 for coefficients and

confidence intervals for all terms) accounted for 10.5% of the attractiveness difference in criterion bias.

Discussion

Raising awareness that applicants will differ on physical attractiveness reduced judgment biases favoring physically attractive people, and debiased behavior among those wanting to treat targets equally or believed they had done so. All interventions were effective only when they raised awareness about differences in applicants' physical attractiveness. Inducing a sense of accountability, providing concrete if-then plans to focus on only task-relevant criteria, and creating commitment to avoid irrelevant information did nothing to reduce attractiveness bias if participants were not alerted to attractiveness of the candidates.

This startling result provides additional evidence that theoretically-diverse interventions for reducing biased judgment may operate through a shared mechanism of raising awareness. In this paradigm, the present evidence suggests that awareness is necessary for interventions to reduce judgment bias, and it is suggestive that awareness may be sufficient to account for the reduction in judgment biases observed in other bias reduction interventions. It is still possible that the unique features of each intervention would be effective at reducing biased behavior with a more powerful test or with a stronger form of social bias. Moreover, there may be particular features of this paradigm, decision context, or social domain enabling awareness to be sufficient on its own. Changes on those dimensions could reveal additional value for theoretically specific elements of each intervention. But, for now, the possibility that the theoretically-unique features have effectiveness above-and-beyond just raising awareness in some circumstances is just speculation.

Study 4a and 4b

Studies 1-3 illustrated the effectiveness of raising awareness in reducing biased judgment, and Study 3 suggests that distinct bias reduction interventions may operate via a shared mechanism of increasing participants' awareness of potential biases. To further investigate how awareness reduces judgment bias, in two final studies, I tested whether creating awareness for one social bias can impact other social biases simultaneously.

In Study 4a and 4b, participants completed an academic JBT in which applicants were presented with relevant academic qualification and two forms of irrelevant but potentially biasing social information. Specifically, in Study 4a, applicants were either more or less physically attractive and came either from one's own or a rival university. In Study 4b, applicants were either more or less physically attractive and belonged to either one's own or another political party. In each study, I manipulated awareness of bias for only one of the two social dimensions (favoring one's own university in Study 4a or favoring one's own political party in Study 4b).

By only targeting one of the potential biases in behavior, these studies can clarify whether increased awareness about a specific bias creates a more global change in social judgment in which awareness of one bias leads participants to be less impacted by *all* irrelevant social information, or if it creates a more local change in which awareness of one bias leads to participants only being less impacted by the targeted bias and other potential biases in behavior are unaffected (or potentially exacerbated). Awareness of one potential bias could inhibit multiple social biases in behavior by activating egalitarian goals (Moskowitz, Gollwitzer, Wasel & Schaal, 1999) or increasing a general motivation to control prejudiced responses (Dunton & Fazio, 1997). Alternatively, awareness of potential bias could be limited to only the identified social dimension if the mechanisms responsible for debiasing require awareness of the bias

target. The results of Study 3 suggest this may be the case, as participants needed to be aware of the specific social dimension impacting judgment in order to reduce bias. In Study 4, I directly test whether awareness of one specific bias in behavior impacts other social biases operating simultaneously.

Method

Participants

Participants in Study 4a were undergraduates who completed the study for either a gift card or partial course credit. We originally targeted a sample of 652 participants. This sample would provide greater than 80% at detecting a between-subjects effect size of $d = .22$, which was the size of the smaller criterion bias found in a pilot study with the modified JBT used in Study 4a.[7] However, since results from the initial sample were inconclusive, we decided to collect as much data as possible during the Spring 2017 semester, again reporting *p*-augmented to account for additional data collection after observing the outcomes. 929 participants have completed Study 4a ($M_{Age} = 18.9$, $SD = 1.2$, 62.5% female, 59.4% White). See https://osf.io/bm5yk/ for pre-registration of materials and original sample size, https://osf.io/r4xvk/ for pre-registration of the analysis plan, and https://osf.io/dpfyv/ for pre-registration of the final data collection strategy.

Participants in Study 4b completed the study through an online survey company. We planned to recruit at least 1200 participants who completed all study measures, which would provide greater than 93% at detecting a small between-subjects effect size of $d = .20$. In the final sample, 1223 participants provided data and passed the attention check measures ($M_{Age} = 42.4$, $SD = 13.0$, 72.6% female, 63.3% White). In both studies, sample sizes vary across tests due to missing data. See https://osf.io/2dpmx/ for the study's pre-registration.

Procedure

[7] See the online supplement for the full report of this pilot study.

Participants in Study 4b completed four study components in the following order: participants first received the bias intervention, then completed the academic JBT, followed by items measuring perceptions of JBT performance, and then measures of explicit attitudes. Participants in Study 4a completed those same components as well as 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 assigned to either a *Control* or *Awareness* condition. Participants in the *Control* condition did not receive any additional instructions. Participants in the *Awareness* condition received the same intervention used in Study 1 except the manipulation now targeted a different social bias. In Study 4a, participants were made aware of a bias favoring applicants from one's own university. In Study 4b, participants were made aware of a bias favoring applicants from one's own political party (see Appendix C for full text of both interventions). Since data collection for Study 4a began before Study 2, the Awareness manipulation included both the awareness and fairness components included used in Study 1. These two components were also included in the awareness manipulation for Study 4b to maximize similarity between conditions.

Academic decision-making task. Participants completed a modified version of the academic JBT. In this version, the same profiles used in Studies 1-3 were shown, now randomizing two pieces of social information. In Study 4a, applicants varied in physical attractiveness (the same faces from Studies 1-3) and school, depicted with a logo of the University of Virginia (UVA), or a rival school, the University of North Carolina (UNC). Instructions stated that both UVA and UNC are equally rigorous, so academic qualifications from both schools should be given the same weight in decision making. In Study 4b, applicants

varied in physical attractiveness and political identity, depicted by a political logo of the Democratic or Republican parties. These logos were a red, white and blue donkey (for Democrats) or elephant (for Republicans).[8] Participants were reminded of the affiliations for each logo before starting the JBT.

The JBT in both studies lasted for 64 trials, with eight trials (four male, four female) for each combination of qualification, physical attractiveness, and secondary social category (e.g., eight more physically attractive and more qualified Democrats, eight less physically attractive and more qualified Democrats). Before evaluating applicants, participants in Study 4a first completed an encoding phase where each applicant was passively shown for one second in a random order. This encoding phase was removed from Study 4b to save time.

The lab participants in Study 4a were assigned to one of two JBT orders. In each order, the specific faces paired with either UVA or UNC were predetermined, but the face-school pairings were then randomly assigned to applications each time the program was run. Across both orders, each application was equally likely to be assigned to either a more vs. less physically attractive face and equally likely to be depicted as coming from UVA vs. UNC. The online participants in Study 4b were assigned to one of twelve study orders, with each application being equally likely to be assigned to a more vs. less physically attractive face or to be depicted as a Democrat or Republican across orders.

Awareness of differences among applicants. Following the JBT, participants in Study 4a used a five-point scale (1= "Not different at all", 5 = "Extremely different") to rate how different applicants on the JBT were on five dimensions: university affiliation, gender, race,

[8] We chose these secondary social categories due to their relevance to the sample population and previous evidence that they impacted performance on the JBT. In Axt et al. (2017), UVA students had lower criterion for a UVA vs. UNC applicant (Study 2a; d = .41), and online participants had lower criterion for members of their own political party compared to the other political party (Study 3; d = .31).

physical attractiveness and facial expression. These items were not in our analysis plan but added for exploratory purposes to see whether the awareness manipulation changed subjective feelings of awareness of differences in applicants on several dimensions.

Perceptions of performance and explicit attitudes. Participants then completed the same measures of perceived performance and desired performance as in Studies 1-3. In Study 4a, participants completed measures separately for performance towards more vs. less physically attractive applicants and for UVA vs. UNC applicants. In Study 4b, participants completed measures separately for performance towards more vs. less physically attractive applicants and for Democrat vs. Republican applicants. In both studies, participants also completed the same explicit preference item used in Studies 1-3 for the two social dimensions used in the JBT.

Implicit evaluations. Participants in Study 4a then completed two seven-block evaluative IATs measuring evaluations towards more vs. less physically attractive people and towards UVA vs. UNC. See Appendix D for more information on the IAT procedure and task stimuli. The two IATs were completed in a random order. IATs were scored with the *D* algorithm (Greenwald, Nosek & Banaji, 2003) such that more positive scores reflected more positive evaluations of more vs. less physically attractive people and of UVA vs. UNC. For analyses involving the IAT, we also excluded participants who had more than 10% of trials faster than 300 milliseconds (.1% of the attractiveness and .5% of school *D* scores; Nosek, Greenwald, & Banaji, 2007).

Demographics. Finally, participants in Study 4a completed a five-item demographics survey, of which we analyzed data concerning age, gender and race. Participants in Study 4b 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 an option other than Democrat or Republican, then they were presented with a forced-choice item asking if they *had* to choose, whether they identified more with Democrats or Republicans.

For analysis, we combined all participants who either initially selected Democrats (or Republicans) or selected Democrats (or Republicans) in the forced-choice item. This maximized power and aligned with prior evidence suggesting that these groups have similar political attitudes and behaviors (e.g., Hawkins & Nosek, 2012). In addition, to further maximize power and simplify analyses, we analyzed data in terms of whether evaluations were made towards applicants from one's own vs. the other political party.

Results

Participants were excluded from analysis for accepting less than 20% or more than 80% of the applicants, or for accepting or rejecting every applicant from any of the social groups used in the JBT. Fifteen participants (1.6%) were excluded based on these criteria in Study 4a and 286 participants (24.1%) were excluded based on these criteria in Study 4b.[9]

In both studies, among eligible participants, overall accuracy on the task was above chance (Study 4a: $M = 70.1\%$, $SD = 7.0$; Study 4b: $M = 62.2\%$, $SD = 9.7$) and the average acceptance rate was close to the requested 50% (Study 4a: $M = 52.9\%$, $SD = 10.1$; Study 4b: $M = 53.4\%$, $SD = 13.6$).

Criterion Bias in Decision-Making

For Study 4a, the primary analysis was a 2 (Attractiveness: More vs. Less physically attractive) by 2 (School: UVA vs. UNC) by 2 (Condition: Awareness vs. Control) repeated

[9] As in Studies 2-3, the higher exclusion rate in Study 4b is likely due to removing the encoding phase. Study 4a was also a lab study, which further limited participants' willingness to drop out. Analyses in the main text use our pre-registered exclusion criteria, but primary conclusions are not altered when including all participants (analyses available in the online supplement).

measures ANOVA on the criterion value 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][10], and school, $F(1, 912) = 8.71$ $p = .003$, $\eta^2_p = .009$, 95% C.I. [.001, .03], $p_{augmented} = [.05, .0502]$. These main effects were in the anticipated directions, with lower criterion for more vs. less physically attractive applicants and for applicants coming from UVA vs. UNC. There was not a reliable main effect of Condition, $F(1, 912) = .42$, $p = .516$, $\eta^2_p < .001$. See Figure 4 for a graphical display of results for Studies 4a and 4b.

Of primary interest were the Attractiveness by Condition and School by Condition interactions. A reliable school by condition would show that making people aware of potential school-affiliation bias reduced school bias, consistent with the prior studies. The school by condition interaction pattern was consistent with this prediction but was not significant, $F(1, 912) = 3.03$ $p = .082$, $\eta^2_p = .003$, 95% C.I. [0, .02], $p_{augmented} = [.082, .119]$. Follow-up ANOVAs showed that the main effect of school was larger in the Control ($F(1, 424) = 9.25$ $p = .002$, $\eta^2_p = .021$, 95% C.I. [.003, .06], $p_{augmented} = [.05, .0501]$) than the Awareness ($F(1, 488) = .87$, $p = .351$, $\eta^2_p = .002$) condition. A reliable attractiveness by condition interaction would show that making people aware of potential *school-affiliation* bias reduced *attractiveness* bias, extending prior findings to suggest that raising awareness of one bias can reduce others. The attractiveness by condition interaction was small but reliable, $F(1, 912) = 5.13$, $p = .024$, $\eta^2_p = .006$, 95% C.I. [0, .02], $p_{augmented} = [.051, .058]$. Follow-up ANOVAs in each condition showed that the main effect of attractiveness was larger in the Control ($F(1, 424) = 80.37$, $p < .001$, $\eta^2_p = .159$, 95% C.I. [.10, .22]) than the Awareness ($F(1, 488) = 48.96$, $p < .001$, $\eta^2_p = .091$, 95% C.I. [.05, .14])

[10] The first sample yielded a highly reliable main effect of awareness ($p = 2.26 \times 10^{-23}$), making the *p*-augmented range very close to .05 and not worth reporting in full.

condition.[11] Awareness of a bias favoring applicants from one's own school had a small effect on attenuating the degree to which participants favored more over less physically attractive applicants.

Unrelated to the key hypotheses, neither the school by attractiveness interaction, $F(1, 912) = 3.72$, $p = .054$, $\eta^2_p = .004$, or the school by attractiveness by condition interaction, $F(1, 912) = 2.23$, $p = .136$, $\eta^2_p = .002$, were reliable. See Table 11 for output from the 2 (Attractiveness) by 2 (School) ANOVAs within in each condition.

For Study 4b, 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) repeated measures ANOVA on the criterion value for applicants from each combination of political party and physical attractiveness. This analysis revealed main effects of physical attractiveness, $F(1, 898) = 105.77$, $p < .001$, $\eta^2_p = .105$, 95% C.I. [.07, .14], and political ingroup, $F(1, 898) = 65.82$ $p < .001$, $\eta^2_p = .068$, 95% C.I. [.04, .10], in the anticipated direction. Participants showed a lower criterion for more vs. less physically attractive applicants and for applicants from one's own vs. the other political party. There was not a reliable main effect of Condition, $F(1, 898) = 0.59$ $p = .441$, $\eta^2_p = .001$.

Of primary interest were the Attractiveness by Condition and Political Group by Condition interactions. A reliable political group by condition would show that making people aware of potential political-affiliation bias reduced political bias. The political group by condition interaction was reliable, $F(1, 898) = 12.19$, $p = .001$, $\eta^2_p = .013$, 95% C.I. [.003, .03]. Follow-up ANOVAs in each condition showed that the main effect of school was larger in the Control ($F(1, 466) = 55.78$, $p < .001$, $\eta^2_p = .107$, 95% C.I. [.06, .16]) than the Awareness ($F(1,$

[11]As in the analysis in the full sample, the main effect of attractiveness was highly reliable in the first sample (Control $p = 1.26 \times 10^{-14}$; Awareness $p = 5.26 \times 10^{-10}$), making the *p*-augmented range very close to .05 and not worth reporting in full.

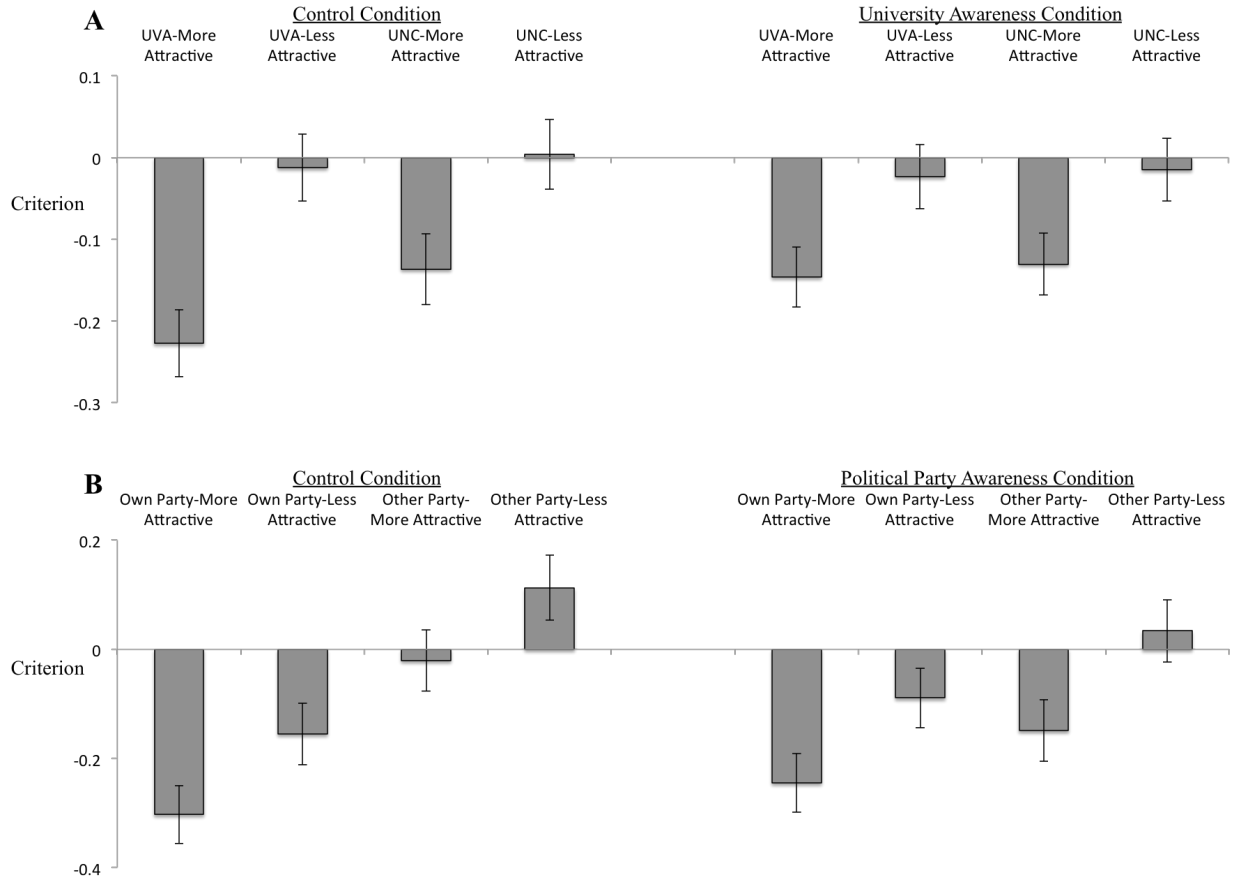

Figure 4. Criterion values for each Control and Awareness condition for Study 4a (Panel A) and Study 4b (Panel B). Lower values mean a lower criterion for applicants from that group. Error bars denote 95% confidence intervals on the mean.

432) = 14.18, $p < .001$, $\eta^2_p = .032$, 95% C.I. [.01, 07]) condition. A reliable attractiveness by condition interaction would show that making people aware of potential *political-affiliation* bias reduced *attractiveness* bias. Unlike Study 4a, the awareness by condition interaction was not reliable, $F(1, 898) = 0.93$, $p = .336$, $\eta^2_p = .001$. Follow-up ANOVAs within each condition showed that the main effect of attractiveness in the Control condition ($F(1, 466) = 44.65$ $p < .001$, $\eta^2_p = .087$, 95% C.I. [.04, .14]) was, if anything, slightly smaller than the main effect in the Awareness condition ($F(1, 432) = 61.70$, $p < .001$, $\eta^2_p = .125$, 95% C.I. [.07, .18]). There was no evidence that awareness of a bias favoring applicants from one's own political group impacted the degree to which participants favored more over less physically attractive applicants.

Unrelated to the key hypotheses, neither the political group by attractiveness interaction, $F(1, 898) = .10$, $p = .755$, $\eta^2_p < .001$, or the political group by attractiveness by condition interaction, $F(1, 898) = .97$, $p = .325$, $\eta^2_p = .001$, were reliable. See Table 11 for output from the 2 (Attractiveness) by 2 (Political Group) ANOVAs within in each condition.

Meta-Analysis of Impact of Awareness on Criterion Bias

The influence of the awareness manipulation across Studies 4a and 4b yielded somewhat different results. In Study 4a, there were very small effects of awareness reducing the targeted bias (favoring of one's own school) and of reducing the unmentioned bias (favoring more physically attractive people). In Study 4b, there was only an effect of awareness reducing the targeted bias (favoring of one's own political party) and no effect of this awareness impacting the other bias (favoring more physically attractive people). To provide more precise estimates of the impact of awareness on both the targeted and untargeted social biases, we conducted a "mini meta-analysis" of Studies 4a and 4b using the tools provided by Goh, Hall & Rosenthal (2016), first converting each partial eta-squared (η^2_p) into a Pearson's correlation (r).

There was a small but reliable meta-analytic effect such that raising awareness of a bias toward a group was associated with a reduced bias toward that group ($r_{\text{study 4a}} = .06$, $r_{\text{study 4b}} = .12$, $r_{\text{meta-analysis}} = .09$, 95% C.I [.03, .15], $Z = 2.96$, $p = .003$). However, there was no reliable meta-analytic effect of raising awareness of bias toward one group on reducing bias toward another, unmentioned group ($r_{\text{study 4a}} = .08$, $r_{\text{study 4b}} = -.03$, $r_{\text{meta-analysis}} = .02$, 95% C.I [-.09, .13], $Z = 0.40$, $p = .687$). See Figure 5 for forest plots of both meta-analyses.

Attitudes, Desired and Perceived Performance

In Study 4a, explicit ($M = 1.04$, $SD = .84$, $d = 1.24$) and implicit ($M = .74$, $SD = .33$, $d = 2.24$) attitudes indicated preference for more physically attractive people. Explicit ($M = .69$, SD

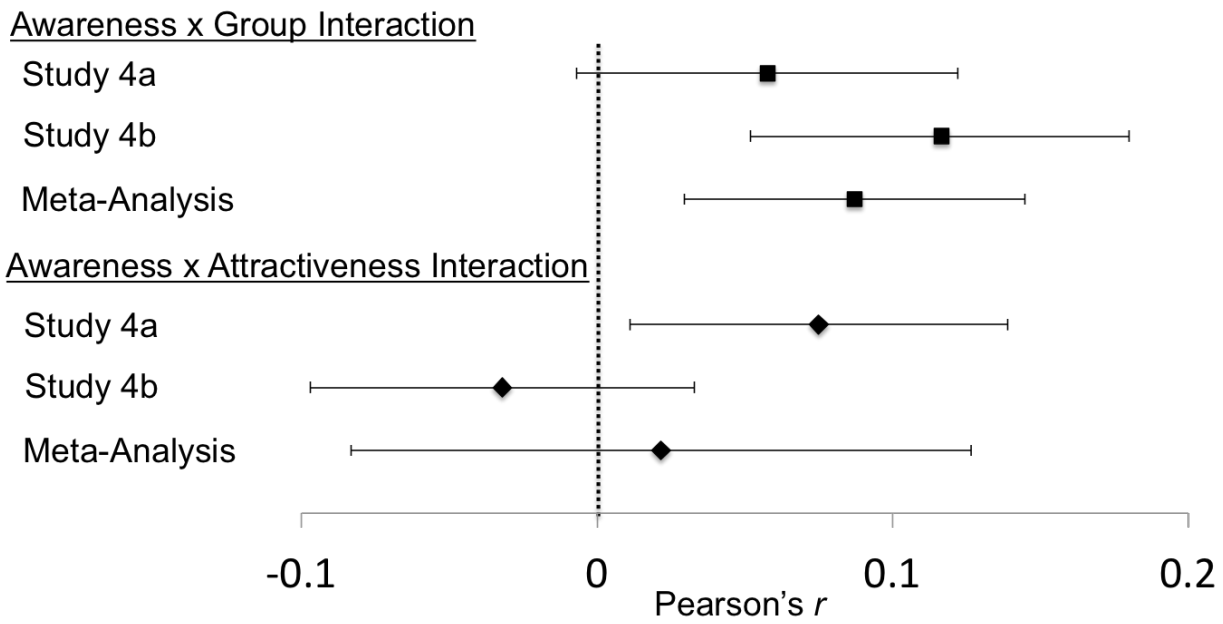

Figure 5. Forest plot for Awareness by Group and Awareness by Attractiveness interactions in Studies 4a and 4b. Error bars denote 95% confidence intervals.

= .90, d = .77) and implicit (M = .40, SD = .31, d = 1.32) attitudes also indicated preference for UVA vs. UNC. In a series of between-subjects t-tests comparing the Control and Awareness conditions on explicit and implicit attitudes, perceived performance and desired performance, the only reliable tests were that participants in the Awareness condition reported lower levels of perceived (M =.10 , SD = .40) and desired (M =.06 , SD = .36) favoritism towards UVA students than participants in the Control (Perceived: M =.19 , SD = .49; Desired: M =.17 , SD = .54) condition (Perceived: $t(912) = 2.98$, $p = .003$, d = .20, 95% CI [.07, .33]; Desired: $t(912) = 3.56$, $p < .001$, d = .24, 95% CI [.11, .37].

In Study 4b, explicit attitudes indicated preference for more physically attractive people (M = 0.45, SD = 1.01, d = 0.44) and for members of one's own political party (M = 1.15, SD = 1.24, d = .93). In a series of between-subjects t-tests comparing the Control and Awareness conditions on explicit attitudes, perceived performance and desired performance, the only

reliable tests were small differences such that participants in the Awareness condition reported lower levels of perceived favoritism towards applicants from their own political party than participants (M = 0.20, SD = 0.73) and lower levels of explicit preference for people from their own political party (M = 1.03, SD = 1.16) than participants in the Control (Perceived: M = .32 , SD = .87; Explicit: M = 1.22 , SD = 1.28) condition (Perceived: $t(885) = 2.22$, $p = .026$, $d = .15$, 95% CI [.02, .28]; Explicit: $t(890) = 2.32$, $p = .021$, $d = .16$, 95% CI [.02, .29]. See Table 12 for means and test statistics for attitudes and performance measures in Studies 4a and 4b.

Predicting Bias in Criterion

As in Studies 2-3**,** we pre-registered analyses that were not central to the study's hypothesis but could be of interest for other purposes. For analyses predicting criterion bias, we only report estimates and confidence intervals. For both studies, we made a criterion difference score for each of the two social biases used in the JBT, re-analyzing the data by collapsing across the other social dimension (e.g. looking only at the criterion for UVA vs. UNC applicants, ignoring whether they were paired with more vs. less physically attractive faces). Higher values indicate a lower criterion for members of one's own school or political group versus the outgroup or for more versus less physically attractive applicants.

In Study 4a, criterion bias for school was positively but weakly related to implicit attitudes (r = .06, 95% CI [-.01, .12]), explicit attitudes (r = .09, 95% CI [.03, .15]), perceived performance (r = .13, 95% CI [.06, .19]), and desired performance (r = .10, 95% CI [.04, .16]). Criterion bias for physical attractiveness was also weakly and positively related to related to implicit attitudes (r = .09, 95% CI [.02, .15]), explicit attitudes (r = .17, 95% CI [.10, .23]), perceived performance (r = .27, 95% CI [.20, .32]), and desired performance (r = .13, 95% CI [.06, .19]).

In Study 4b, criterion bias for political party was positively and more strongly related to explicit attitudes (r = .23, 95% CI [.17, .29]), perceived performance (r = .43, 95% CI [.37, .48]), and desired performance (r = .38, 95% CI [.32, .43]). Criterion bias for physical attractiveness was very weakly and positively related to related to explicit attitudes (r = .09, 95% CI [.03, .16]), but negatively and very weakly related to perceived performance (r = -.01, 95% CI [-.07, .06]), and desired performance (r = -.01, 95% CI [-.07, .06]).

For each study and each social dimension, we also ran simultaneous linear regressions predicting criterion bias variable from the relevant attitude and performance measures. In Study 4a, these variables explained 2.8% of the variance in school criterion bias and 8.1% of the variance in attractiveness criterion bias. In Study 4b, these variables explained 21.3% of the variance in political party criterion bias and 0.9% of the variance in attractiveness criterion bias. See Table 13 for regression estimates and confidence intervals for each term. Finally, we ran a simultaneous linear regression for each criterion bias, now including condition (coded with Control as the reference) and interactions with each relevant attitude and performance measure. In Study 4a, these variables explained 3.7% of the variance in school criterion bias and 9.8% of the variance in attractiveness criterion bias. In Study 4b, these variables explained 26.0% of the variance in political party criterion bias and 1.5% of the variance in attractiveness criterion bias. See Table 14 for regression estimates and confidence intervals.

Discussion

Studies 4a and 4b extend the earlier studies by illustrating that raising awareness about political or school-affiliation (ingroup or outgroup) biases can reduce biased judgment. This is notable in that these social differences were quite obvious (e.g., logos clearly depicting school or

political affiliation), and perhaps easier to spontaneously recognized as a source of potential bias compared to physical attractiveness.

In Study 4a, there was a small effect such that awareness of a bias favoring university ingroup members attenuated a second, unmentioned bias of favoring more physically attractive people. A similar effect did not emerge in Study 4b; there was no evidence that awareness of a bias favoring political ingroup members impacted bias towards more vs. less physically attractive people. It is possible that differences in samples or the social domains used in the JBT account for these contrasting findings. However, given the high degree of similarity between the two studies, we believe the most precise estimates come from the meta-analytic effects across both studies, which found that awareness reduced the targeted but not the unmentioned bias.

From this perspective, results do not support the account that increased awareness of one bias can impact other biases operating simultaneously, such as through making egalitarian goals more accessible (Moskowitz, Salomon & Taylor, 2000) or increasing motivation to control biased behavior (e.g., Maddux, Barden, Brewer & Petty, 2005). The aggregate results of Studies 4a and 4b align with the results of Study 3 in which interventions that did not identify the specific bias were ineffective at reducing biased judgment.

Interestingly, Study 4a showed small but reliable effects of inducing awareness on reducing perceived and desired bias favoring applicants from one's own university, and Study 4b found similar effects for perceived but not desired bias regarding treatment of applicants from one's own versus another political group. These results are inconsistent with Studies 1-3 in which awareness interventions did not consistently influence perceived or desired performance. It is possible that awareness interventions have more of an impact on desired and perceived performance towards the targeted social bias in the context of the simultaneous JBT used in

Studies 4a-4b, though given the small effects found here, additional data are likely needed to clarify in what samples and domains awareness impacts perceptions of fairness and desires for fair treatment.

Gender Bias and Awareness Interventions

In all five experiments, applicant gender was manipulated but ignored. Half of the more qualified and less qualified candidates in every study were female and half were male, and these were crossed with the other manipulated variables – attractiveness, politics, and university affiliation. For exploratory purposes, I analyzed gender as another potential bias for all five studies.

In a prior investigation using this attractiveness JBT (Axt et al., 2017, Study 1a), criterion bias differences were slightly favoring female compared to male candidates ($N = 203$, $d = .20$, $p = .131$). This effect may have been reasonably accurately estimated but underpowered. This appears to be the case. With larger samples, we observed a reliable criterion bias favoring females over males . In the control condition female applicants received a lower criterion than male applicants in all but one study (Study 1: $d = .25$, $p < .001$; Study 2: $d = .17$, $p = .011$; Study 3: $d = .31$, $p < .001$; Study 4a : $d = .003$, $p < .948$; Study 4b: $d = .38$, $p < .001$). A meta-analysis combining control conditions for all five studies show a reliable difference in criterion bias for gender ($d = .23$, $p < .001$). Likewise, considering only the experimental conditions of all five studies, there was a reliable difference in criterion bias for gender ($d = .21$, $p < .001$). See the online supplement for full analyses from each study.

Applicant gender was never mentioned in any experiment. This presents an opportunity to test the purpose of Studies 4a and 4b in every experiment. Did making participants aware of potential for attractiveness bias have an impact gender bias? Across studies where participants

showed a gender bias in Control conditions, comparing Awareness and Control conditions did not reveal a change in criterion bias for gender in all but one study (Study 1: $d = .11$, $p = .042$; Study 2: $d = .02$, $p = .800$; Study 3: $d = .05$, $p = .474$; Study 4b: $d = .08$, $p = .245$).[12] These results suggest a similar conclusion to the meta-analytic estimates of Studies 4a and 4b on attractiveness biases; making participants aware of one bias did not consistently influence another bias that was operating simultaneously.

The gender bias observed here may present avenues for future work to identify the unique contribution of the bias reduction strategies beyond raising awareness. There were hints of other effects in exploratory analyses. For instance, in Study 1, there was some evidence that implementation intentions ($d = .16$, $p = .004$) reliably reduced gender bias relative to Control. Similarly, in Study 3, the awareness-raising versions of each intervention reduced gender bias relative to Control (Accountability: $d = .13$, $p = .060$; Implementation Intentions: $d = .14$, $p = .038$; Objectivity: $d = .14$, $p = .035$). See the online supplement for analyses.

General Discussion

Across five studies, participants made aware of a specific bias in social judgment showed reduced bias towards targets from those social groups (more vs. less physically attractive people in Studies 1-3, university or political ingroup vs. outgroup members in Studies 4a-4b). Previously identified bias reduction strategies attenuated biased judgment, but these interventions were only effective when including text that made participants aware of the specific social dimension that could be the source of bias. Raising awareness of possible bias was more important for reducing judgment biases than asking participants to behave fairly.

[12] For Study 1, this analysis compared the Awareness and Control condition. For Study 2, this analysis is the main effect of Awareness on gender bias in a 2 (Awareness) by 2 (Fairness) ANOVA. In Study 3, this analysis compared the Awareness and No Awareness versions of the awareness intervention. Study 4a was not included because the Control condition did not show a reliable gender bias in criterion.

This work replicates and extends prior research on reducing biased behavior, showing that interventions such as creating implementation intentions (Mendoza, Gollwitzer & Amodio, 2010; Stewart & Payne, 2008), increasing accountability (Webster, Richter & Kruglanski, 1996), and highlighting inconsistencies between values and actions (Monteith & Voils, 1998) all reduce socially biased judgment in a new task (the JBT), social domain (physical attractiveness), and decision context (making academic evaluations). More importantly, by comparing these interventions simultaneously, the findings suggest that they may operate via a shared mechanism of increasing awareness. This possibility has dramatic implications for the theoretical interpretation of these interventions.

Implementation intentions are believed to reduce biased judgment by limiting the influence of automatic associations (Mendoza et al., 2010), or facilitating goal-directed behavior (Stewart & Payne, 2008). Activating should-would discrepancies is believed to reduce biased judgment by creating negative self-directed affect (Monteith, 1993) and increasing motivations to inhibit bias in the future (Monteith, Mark & Ashburn-Nardo, 2010). And, accountability interventions are believed to reduce bias by prompting more critical thinking (Lerner & Tetlock, 1999), minimizing reliance on surface cues (Bodenhausen, Kramer & Susser, 1994) and widening the amount of information used when making decisions (Webster, Richter & Kruglanski, 1996).

These theoretical explanations identify distinct mechanisms for reducing judgment bias, and they co-exist as independent areas of inquiry. If, in fact, all these interventions are actually effective because of a common mechanism of awareness, there are two important implications: (1) intervening on judgment biases may be more simple than previously understood; and (2) there may be no theoretical or practical value for the enriched theoretical explanations of the

unique qualities of each intervention. The strongest evidence for this claim comes from Study 3. Here, each of the interventions that reduced bias on the JBT in Study 3 were ineffective when changing only a few words that either did or did not raise awareness that targets would differ in physical attractiveness (e.g., changing "physical attractiveness" to "irrelevant information"). Moreover, as seen in both Study 1 and Study 3, even those interventions that did raise awareness were no more effective than an intervention that simply warned participants about a possible bias in evaluation.

At most, these data indicate that heightened awareness of a specific bias in social judgment is sufficient to attenuate biased behavior on that social dimension, and additional characteristics unique to each intervention beyond raising awareness do not increase effectiveness. However, it is possible that the bias reduction strategies tested here may show additional effectiveness beyond raising awareness when applied to other types of judgment or towards different social groups. For instance, the effectiveness of raising awareness could be limited to judgments that are relatively easy to correct; other strategies may show added value when applied to behaviors that are more difficult to control, such as those made with time pressure (e.g., Mendoza et al., 2010) or under cognitive load. For example, there is some evidence that implementation intentions can operate efficiently or even automatically (Brandstatter, Lengfelder & Gollwitzer, 2001; Gollwitzer & Schaal, 1998). Drawing from this work, it would be useful to add a manipulation of time pressure to the present paradigm to test whether awareness itself continues to be sufficient, or if it declines while the effectiveness of the implementation intentions intervention is preserved.

Another possible constraint on the generalizability of awareness as the intervention mechanism is that over 85% of participants in these studies reported a desire to behave fairly.

The interventions studied here may be more effective than awareness alone when targeting biases that people do not necessarily oppose, such as towards drunk drivers or perpetrators of domestic violence (Crandall, Eshleman & O'Brien, 2002). In existing models of bias correction, motivation to overcome bias is perceived as an important element for bias reduction (Fazio, 1990; Wegener & Petty, 1996; Wilson & Brekke, 1994). However, an exploratory meta-analysis across Studies 1-3 suggests that awareness reduced criterion bias even among participants *not* wanting to treat applicants equally (Study 1: $r = .20$, Study 2: $r = .14$, Study 3: $r = .11$; Meta-Analysis $r = .13$, 95% CI [.05, .21]). Also, we did include a common, simple motivation induction in Study 2 asking participants to "be fair towards all applicants" and that had no impact on criterion bias. Even so, more direct testing of the role of motivation is needed in this paradigm and crossed with manipulations of awareness. It would be quite surprising if there were no independent role of motivation in reducing bias expression, particularly considering the existing theory and evidence (e.g., Devine et al., 2002; Crandall & Eshleman, 2003). Even in the case of discovering a role of motivation, it will still be of interest to understand whether these theoretically-rich interventions are actually distinct in engaging such motivations.

Additional investigations on the impact of these bias reduction strategies could assess whether similar results emerge when tested across different measures of socially biased behavior. For instance, it is possible that these interventions may show additional effectiveness beyond awareness when applied to less controlled behaviors than those in the JBT, such as in seating distance (e.g., Zogmaister, Arcuri, Castelli, & Smith, 2008) or nonverbal cues like eye contact (e.g., Dasgupta & Rivera, 2006) during interactions with outgroup members. These bias reduction strategies may also show effectiveness beyond awareness when using other forms of each intervention, such as by activating desires for objective behavior through should-would

discrepancies (Monteith & Devine, 1993), strengthening feelings of accountability by having participants believe their responses will be viewed by their peers (Hess, Rosenberg, & Waters, 2001), or creating implementation intentions that explicitly ask participants to be unbiased (e.g., Webb, Sheeran, & Pepper, 2012) rather than to simply ignore "irrelevant information" as in Study 3.

Finally, the bias reduction strategies used here (increasing accountability, committing to objective behavior, creating implementation intentions) are not a random or representative sample of interventions to combat socially biased behavior. Other approaches to lessen biased judgment, such as changing attitudes towards targets through evaluative conditioning (Kawakami et al., 2008), providing more individuating or contextual information (Malinen, Willis & Johnston, 2014; Neuberg & Fiske, 1987), or simulating contact with outgroup members (Todd, Bodenhausen, Richeson & Galinsky, 2011), may be less dependent on raising awareness of bias. The present investigation offers a framework for comparative testing of bias reduction strategies and identification of their shared and unique theoretical mechanisms. A meta-theory accounting of the mechanisms contributing to bias reduction may yield insights for novel interventions that are even more effective.

Awareness as a simple and effective intervention

This work further establishes the causal effect of increased awareness on reducing biased judgment, complementing previous correlational (Pope, Price & Wolfers, 2016) and experimental studies (Golding, Fowler, Long, & Latta, 1990; Petty & Cacioppo, 1986; Schul, 1993). The results presented here show that raising awareness can effectively attenuate socially biased judgment even in domains (like physical attractiveness) where bias often occurs outside

of conscious awareness or control (Axt, et al., 2017). Awareness interventions that target a specific social bias appear to be an efficient and flexible approach for reducing biased behavior.

The generalizability of these awareness interventions is still to be determined. For one, awareness interventions may be more effective for some social categories than others. In these studies, the biasing influence of physical attractiveness or shared university and political affiliations may not be obvious, and efforts to correct bias may not occur spontaneously. For such categories, awareness of bias may be needed to instill the motivation and ability to attenuate bias in judgment. For other social categories (e.g., gender and race), awareness to avoid bias may occur spontaneously, making an additional awareness intervention ineffective. In fact, people may sometimes be so attuned to possible bias that they overcorrect their behavior. White participants completing an academic JBT with White and Black faces actually had lower criterion for Black than White applicants, despite implicit and explicit attitudes that favored White people (Axt, in press; Axt, et al., 2016). It is also possible that awareness could ironically strengthen bias in some circumstances, such as if an awareness intervention identified a social category that would not have otherwise influenced judgment. By drawing attention to the category, decision-makers may be more likely to be influenced by it.

The durability of awareness interventions is also unknown. In the present studies, the judgment task occurred during the six minutes immediately following the intervention. It is easy to anticipate that raising awareness is effective on a limited time scale, perhaps as brief as minutes or hours (e.g., Lai et al., 2016). Time delay may be another opportunity to identify unique contributions of the theoretically-rich aspects of the interventions evaluated here. It is possible, for example, that establishing an implementation intervention is effective in the short-

term because of raising awareness, but in the long-term because of the sticking power of the intention (Holland, Aarts & Langendam, 2006; Gollwitzer & Brandstatter, 1997).

Finally, future studies will need to clarify why the awareness manipulation used here reduced biased judgment in light of previous studies finding that messages conveying high prevalence of bias can *increase* prejudicial attitudes and behavior. For instance, participants told that "the vast majority of people" hold stereotypical preconceptions later expressed more explicit age and weight stereotypes (Duguid & Thomas-Hunt, 2015). Follow-up studies showed that similar messages resulted in more stereotypically biased evaluations and behavior towards a female job applicant. In other work, participants asked to think of the importance of controlling prejudiced behavior, by responding to items like "it is socially unacceptable to discriminate based on cultural background", demonstrated more explicit and implicit racial prejudice (Legault, Gutsell & Inzlicht, 2011).

Some of the manipulations used here had similar components, such as the Awareness manipulation noting that people are frequently easier on more physically attractive and tougher on less physically attractive applicants, or the Commitment manipulation asking participants to indicate beforehand whether it is ever appropriate to use physical attractiveness when making admissions decisions. Despite the parallels to earlier work, these manipulations resulted in reduced biased behavior and no consistent change in explicit or implicit attitudes. One possibility is that participants in the present work interpreted the intervention more seriously and were more likely to believe that they were just as affected as others by the biases brought to their attention. In previous studies, which used domains like gender and race, participants may have been able to view themselves as less prejudiced than others, creating greater license to express prejudice or make biased judgments. In the domains studied here, participants confronted with information

about the prevalence of bias or the importance of confronting it may have had a harder time explaining away that message, creating greater internalization and reduced biases in behavior. One clear method of testing this account would be to apply the same awareness intervention used here to other social domains.

Finally, the awareness manipulation used here simultaneously alerted participants to the social dimension in which applicants differed and indicated that there was an existing bias favoring members of a certain group. These two components—awareness of the relevant social dimension vs. awareness that a bias exists-- were combined in the present studies to maximize the impact of awareness. Future work should investigate whether one or both of these components are necessary to reduce biased judgment, and whether the effectiveness of each component differs across social domains. For example, many participants may spontaneously infer that there is a bias favoring more physically attractive people, so alerting them to this bias does not strengthen the impact of simply being aware that targets differ in attractiveness. However, in domains where people may not have a strong inference about the direction of bias, awareness of the direction of bias may be a necessary component for reducing biased judgment.

The mechanism's mechanism

How does awareness attenuates biased judgment? The results from Studies 4a and 4b suggest that awareness influences processes specific to the targeted social bias. This is an important constraint to identify, but it provides little insight on the specific psychological changes that reduce the identified social bias.

Awareness could impact multiple processes, and there are several plausible means by which awareness reduces biased judgment. For instance, according to Wilson and Brekke's (1994) model awareness is an initial requirement that allows for other psychological changes

necessary to eliminate bias. Specifically, greater awareness of the presence and direction of bias in behavior can then alter 1) motivation to correct bias and 2) ability to control biased responses.

Awareness increasing motivation to be unbiased would align with previous work on mechanisms behind bias reduction interventions. In one study, participants who reflected on a time that they acted toward a Black person in a manner they later regretted were very likely to express a desire to better self-regulate prejudiced thoughts and actions in the future (Monteith, Mark & Ashburn-Nardo, 2010). In other work, participants high in motivation to control prejudice expressed greater self-directed negative affect (e.g., shame or embarrassment) when given false feedback indicating they had shown high levels of implicit bias (Fehr & Sassenberg, 2010). Finally, an intensive six-week bias reduction intervention that raised awareness of bias (among other strategies) increased concerns about discriminatory behavior (Devine, Forscher, Austin & Cox, 2012; Forscher et al., 2017).

A similar process could have occurred among participants in these studies made aware of potential biases in their own behavior, prompting greater motivation to reduce the impact of biasing social information when completing the JBT. Though this motivational account is intuitive, awareness was not consistently associated with a greater desire to treat applicants more fairly, as only one of the five studies (Study 4a) found that awareness led to desired performance that was more fair. The available data do not suggest that awareness reliably impacted motivations to be unbiased, though it is possible that an effect could have emerged on related measures (e.g., concerns over bias, feelings of guilt over one's own behavior).

Awareness could have also improved the ability to adjust one's behavior, resulting in increased effort and reduced social bias on the JBT. In more basic cognitive research, greater attention towards a task has been associated with enhanced perception (Carrasco, Ling & Read,

2004) and more effort (Reynolds, Pasternak & Desimone, 2000). A similar increase in effort created by the awareness intervention could have manifested in two ways. First, greater awareness could have increased the total time spent evaluating applicants as more attention is given to the relevant academic qualifications. However, exploratory analyses across studies using only one social bias (Studies 1-3) found that awareness interventions did not reliably create longer (log-transformed) reaction times on the JBT (r = .03, 95% C.I. [-.02, .09]; see Online Supplement for full details).

Alternatively, awareness could have impacted attentional *allocation*. Participants made aware of a possible bias could have spent the same amount of time on decisions in the JBT, but with a greater percentage of that time focused on the relevant academic qualifications and less on the social information. This perspective would likely predict increased sensitivity (i.e., fewer errors) among participants in Awareness conditions, who presumably spent more time focusing on task-relevant information. Contrary to this expectation, awareness did not reliably impact sensitivity in any single study and an exploratory meta-analysis of Studies 1-3 found no overall effect as well (r = -.001, 95% C.I. [-.04, .04]; see Online Supplement for full details). However, awareness could have increased attention or effort in ways that did not impact overall reaction time or sensitivity, and future studies could use methodologies better suited to monitor attention (e.g. eye-tracking).

The challenge of designing interventions that reduce multiple biases

These studies present multiple avenues to reduce biased judgment within a targeted social domain, but the results of Study 3 and Studies 4a and 4b present an intriguing challenge for developing interventions that can effectively reduce multiple social judgment biases. In Study 3, only interventions that raised awareness of bias in the focal dimension (physical attractiveness)

reduced criterion bias on the JBT, yet in Studies 4a and 4b, there was on average no reliable evidence that similarly targeted awareness interventions can lessen another social bias in behavior, and the analyses concerning gender showed another example of how raising awareness of one bias did not impact multiple biases.

It is possible that awareness of one bias could have motivated participants to be treat all social groups more fairly, and the increased salience of these egalitarian goals could have lessened the impact of pre-existing attitudes or stereotypes on behavior (e.g., Moskowitz & Li, 2011). Or, awareness of one bias could have resulted in participants adopting a strategy of purposefully ignoring all potentially biasing social information and focusing attention only on the relevant judgment criteria (e.g., Goldin & Rouse, 2000). Instead, results from Studies 4a and 4b indicate that awareness only impacted the targeted social bias and left other biases unaffected.

It would be informative to know if participants made aware of multiple social biases (i.e., that people favor both the physically attractive and members of one's own political ingroup) showed reduced bias in both domains, but this approach depends on the capacity to identify potential biases ahead of time. A more effective and generalizable strategy would be to find bias reduction interventions that are broad enough to impact both known and unknown social biases by instilling the motivation or ability needed to focus only on task-relevant information. The present work suggests that developing such interventions will be difficult and highlights a need for more far-reaching bias reduction strategies. The JBT used in Studies 4a and 4b provides one way of measuring simultaneous biases in social judgment, and perhaps the most fruitful next step in addressing this issue is through a "contest" study (e.g., Lai et al. 2014) in which a number of researchers submit and test candidate interventions that they believe will impact multiple social

judgment biases. This approach would accelerate progress on a theoretically and practically important problem.

Limitations

The findings provide many avenues for additional research. Future studies could also address some specific limitations of the methodology in these studies. For one, the online studies had relatively high rates of participant exclusion (Study 1: 9.0%, Study 2 19.1%, Study 3: 15.6%; Study 4b: 24.1%). One cause for the high exclusion rate in Studies 2-4b is the removal of the passive encoding phase. While many of the inattentive participants would have previously dropped out of the study in the encoding phase, removing encoding allowed these participants to complete the JBT but in a manner suggesting they were not following study instructions (e.g., giving the same response to all applicants). Future studies using the JBT can provide more accurate estimates of these effects by reducing both dropout and the percentage of participants excluded from analysis, such as by running more lab studies (where dropout is much lower), or by altering JBT instructions or design to limit the amount of careless responding.

Second, characteristics of the JBT used in these studies may be considered unrepresentative of how many evaluations occur. For instance, in this version of the JBT, participants only viewed each applicant once and were unable to revisit their decisions. Similarly, the even 50%-50% distribution between more and less qualified applicants and between members of the different social categories could impact the process by which participants evaluate applicants and exhibit social bias. Future studies using the JBT should seek to use more diverse methodologies, such as by giving lab participants physical copies of profiles and allowing them to revisit their judgments, or by manipulating the percentages of the applicant

pool that: 1) participants are told to accept, 2) represent more or less qualified applicants, or 3) come from various social groups.

This work will shed light on the robustness of the JBT as a method for studying social bias and highlight those contexts that may alleviate or exacerbate social bias in behavior. It will be informative to see whether the social biases and effective intervention strategies found here are also reproduced in these additional JBT formats. For instance, asking participants to only accept a small proportion of the applicants (e.g., 10%) may increase the effort put into each judgment, thereby reducing both the size of any social biases and the effectiveness of any intervention strategy. Alternatively, when acceptances are thought to be relatively rare, participants may exhibit less vigilance on each individual trial, increasing the degree of social bias. Such an outcome would be similar to findings from the visual search literature showing that errors increase as targets become more infrequent (e.g., Mitroff & Biggs, 2014).

Conclusion

Past research has identified multiple ways to reduce bias in social judgment. In a simultaneous comparison of several prominent interventions, each was capable of reducing judgment biases but only when raising awareness of potential biases within a specific social dimension. Raising awareness may be sufficient to reduce biased judgment, acting as a shared mechanism behind many common strategies. However, the benefits of increased awareness were limited to specifically targeted social biases. Future work in this area will need to further identify the process by which awareness impacts behavior, and use this insight to design interventions that can reduce the impact of multiple biases in social judgment.

References

Ashton, R. H. (1992). Effects of justification and a mechanical aid on judgment performance. *Organizational Behavior and Human Decision Processes*, *52*(2), 292-306.

Axt, J.R. (in press). An unintentional pro-Black bias in judgment among educators. *British Journal of Educational Psychology*.

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), 1-40.

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. University of Virginia.

Bartos, V., Bauer, M., Chytilová, J., & Matějka, F. (2013). Attention discrimination: Theory and field experiments. *CERGE-EI Working Paper Series*, 499.

Beehr, T. A., & Gilmore, D. C. (1982). Applicant attractiveness as a perceived job-relevant variable in selection of management trainees. *Academy of Management Journal*, *25*(3), 607-617.

Bertrand, M., Chugh, D. & Mullainathan, S. (2005). Implicit discrimination. *American Economic Review*, *95*(2), 94-98.

Bertrand, M., & Duflo, E. (2016). Field experiments on discrimination. *National Bureau of Economic Research*, w22014.

Biernat, M., & Fuegen, K. (2001). Shifting standards and the evaluation of competence: Complexity in gender-based judgment and decision making. *Journal of Social Issues*, *57*(4), 707-724.

Bohnet, I., Van Geen, A., & Bazerman, M. (2015). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, *62*(5), 1225-1234.

Bodenhausen, G. V., Kramer, G. P., & Süsser, K. (1994). Happiness and stereotypic thinking in social judgment. *Journal of Personality and Social Psychology*, *66*(4), 621-632.

Brandstätter, V., Lengfelder, A., & Gollwitzer, P. M. (2001). Implementation intentions and efficient action initiation. *Journal of Personality and Social Psychology*, *81*(5), 946-960.

Carrasco, M., Ling, S., & Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, *7*(3), 308-313.

Casey, P.M., Warren, R.K., Cheesman, F.L. & Elek, J.K. (2012). Helping courts address implicit bias: Resources for education. Report prepared for National Center for State Courts.

Cash, T. F., & Kilcullen, R. N. (1985). The Aye of the beholder: Susceptibility to sexism and beautyism in the evaluation of managerial applicants. *Journal of Applied Social Psychology*, *15*(4), 591-605.

Cialdini, R. B. (1984). *Influence: How and why people agree to things*. New York: Quill.

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

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359-378.

Crandall, C. S., & Eshleman, A. (2003). A justification–suppression model of the expression and experience of prejudice. *Psychological Bulletin*, *129*(3), 414-446.

Dasgupta, N., & Rivera, L. M. (2006). From automatic antigay prejudice to behavior: The moderating role of conscious beliefs about gender and behavioral control. *Journal of Personality and Social Psychology*, *91*(2), 268-280.

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., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology*, *60*(6), 817-830.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice. *Journal of Personality and Social Psychology*, *82*(5), 835-848.

Dobbs, M., & Crano, W. D. (2001). Outgroup accountability in the minimal group paradigm: Implications for aversive discrimination and social identity theory. *Personality and Social Psychology Bulletin*, *27*(3), 355-364.

Doleac, J. L., & Stein, L. C. (2013). The visible hand: Race and online market outcomes. *The Economic Journal*, *123*(572), F469-F492.

Duguid, M. M., & Thomas-Hunt, M. C. (2015). Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes. *Journal of Applied Psychology*, *100*(2), 343-359.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*(3), 316-326.

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: Academic Press.

Fehr, J., & Sassenberg, K. (2010). Willing and able: How internal motivation and failure help to overcome prejudice. *Group Processes & Intergroup Relations*, *13*(2), 167-181.

Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin*, *111*(2), 304-341.

Festinger, L. (1957). *A theory of cognitive dissonance*. London, England: Tavistock.

Forscher, P.S., Mitamura, C., Dix, E.L., Cox, W.T.L., & Devine, P.G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. Manuscript submitted for publication. University of Wisconsin at Madison.

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.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, *90*(4), 715-741.

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.

Gollwitzer, P. M., & Brandstätter, V. (1997). Implementation intentions and effective goal

pursuit. *Journal of Personality and Social Psychology*, *73*(1), 186-199.

Gollwitzer, P. M., & Schaal, B. (1998). Metacognition in action: The importance of implementation intentions. *Personality and Social Psychology Review*, *2*(2), 124-136.

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.

Grewal, D., Ku, M. C., Girod, S. C., & Valantine, H. (2013). How to recognize and address unconscious bias. In *The Academic Medicine Handbook* (pp. 405-412). Springer: New York.

Handelsman, J. & Sakraney, N. (2015). Implicit bias. Report prepared for White House Office of Science and Technology Policy

Hansen, K., Gerbasi, M., Todorov, A., Kruse, E., & Pronin, E. (2014). People claim objectivity after knowingly using biased strategies. *Personality and Social Psychology Bulletin*, *40*(6), 691-699.

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.

Hawkins, C.B. Sinden, M. & Nosek, B.A. (2016). Effects of accountability on ingroup favoritism. Manuscript in preparation. University of Illinois at Springfield.

Hess, T. M., Rosenberg, D. C., & Waters, S. J. (2001). Motivation and representational processes in adulthood: The effects of social accountability and information relevance. *Psychology and Aging*, *16*(4), 629-642.

Holland, R. W., Aarts, H., & Langendam, D. (2006). Breaking and creating habits on the working floor: A field-experiment on the power of implementation intentions. *Journal of Experimental Social Psychology*, *42*(6), 776-783.

Hosoda, M., Stone-Romero, E. F., & Coats, G. (2003). The effects of physical attractiveness on

Job-related outcomes: A meta-analysis of experimental studies. *Personnel Psychology*, *56*(2), 431-462.

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors?. *Human Communication Research*, *41*(4), 557-576.

Johnson, V. E., & Kaplan, S. E. (1991). Experimental evidence on the effects of accountability on auditor judgments. *Auditing: A Journal of Practice & Theory*, *10*, 96-107.

Johnson, S. K., Podratz, K. E., Dipboye, R. L., & Gibbons, E. (2010). Physical attractiveness biases in ratings of employment suitability: Tracking down the "beauty is beastly" effect. *The Journal of Social Psychology*, *150*(3), 301-318.

JSExpPlayer [Computer software]. (2015). https://github.com/ProjectImplicit/JSExpPlayer

Kaas, L., & Manger, C. (2012). Ethnic discrimination in Germany's labour market: A field experiment. *German Economic Review*, *13*(1), 1-20.

Kawakami, K., Steele, J. R., Cifa, C., Phills, C. E., & Dovidio, J. F. (2008). Approaching math increases math= me and math= pleasant. *Journal of Experimental Social Psychology*, *44*(3), 818-825.

Kiefer, A., Sekaquaptewa, D., & Barczyk, A. (2006). When appearance concerns make women look bad: Solo status and body image concerns diminish women's academic performance. *Journal of Experimental Social Psychology*, *42*(1), 78-86.

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.

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science*, *22*(12), 1472-1477.

Lelkes, Y. & Westwood, S. (2017). The limits of partisan prejudice. *Journal of Politics*, *79*(2), 485-501.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255-275.

Maddux, W. W., Barden, J., Brewer, M. B., & Petty, R. E. (2005). Saying no to negativity: The effects of context and motivation to control prejudice on automatic evaluative responses. *Journal of Experimental Social Psychology*, *41*(1), 19-35.

Malinen, S., Willis, G. M., & Johnston, L. (2014). Might informative media reporting of sexual offending influence community members' attitudes towards sex offenders?. *Psychology, Crime & Law*, *20*(6), 535-552.

McManus, I. C., Richards, P., Winder, B. C., Sproston, K. A., & Styles, V. (1995). Medical school applicants from ethnic minority groups: Identifying if and when they are disadvantaged. *British Medical Journal*, *310*(6978), 496-500.

Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin*, *36*(4), 512-523.

Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination an audit study in academia. *Psychological Science*, *23*(7), 710-717.

Mitroff, S. R., & Biggs, A. T. (2014). The ultra-rare-item effect visual search for exceedingly rare items is highly susceptible to error. *Psychological Science*, *25*(1), 284-289.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology*, *65*(3), 469-485.

Monteith, M. J., Deneen, N. E., & Tooman, G. D. (1996). The effect of social norm activation on the expression of opinions concerning gay men and Blacks. *Basic and Applied Social Psychology*, *18*(3), 267-288.

Monteith, M. J., Devine, P. G., & Zuwerink, J. R. (1993). Self-directed versus other-directed affect as a consequence of prejudice-related discrepancies. *Journal of Personality and Social Psychology*, *64*(2), 198-210.

Monteith, M. J., Mark, A. Y., & Ashburn-Nardo, L. (2010). The self-regulation of prejudice: Toward understanding its lived character. *Group Processes & Intergroup Relations, 13*(2), 183-200

Monteith, M. J., & Voils, C. I. (1998). Proneness to prejudiced responses: Toward understanding the authenticity of self-reported discrepancies. *Journal of Personality and Social Psychology*, *75*(4), 901-916.

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), 167-184.

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., 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.

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

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.

Munro, G. D., Lasane, T. P., & Leary, S. P. (2010). Political partisan prejudice: Selective distortion and weighting of evaluative categories in college admissions applications. *Journal of Applied Social Psychology*, *40*(9), 2434-2462.

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, *53*(3), 431-444.

Nosek, B. A., Bar-Anan, Y., Sriram, N., Axt, J.R., & Greenwald, A. G. (2014). Understanding and using the Brief Implicit Association Test: Recommended scoring procedures. *PloS ONE*, *9*(12), e110938.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In *Automatic processes in social thinking and behavior* (pp. 265-292). New York: Psychology Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259.

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

937-975.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Communication and Persuasion* (pp. 1-24). Springer: New York.

Pope, D. Price, J. & Wolfers, J. (2016). Awareness reduces racial bias. Manuscript submitted for publication. University of Chicago.

Price, J. & Wolfers, J. (2010). Racial discrimination among NBA referees. *The Quarterly Journal of Economics, 125*(4), 1859-1887

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *43*(4), 565-578.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, *26*(3), 703-714.

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.

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.

Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: The role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*, 583–592

Siegel-Jacobs, K., & Yates, J. F. (1996). Effects of procedural and outcome accountability on judgment quality. *Organizational Behavior and Human Decision Processes*, *65*(1), 1-17.

Spitzer, H., Desimone, R., & Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science*, *240*(4850), 338-340.

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology*, *56*(4), 283-294.

Staats, C., Capatosto, K., Wright, R.A., Contractor, D. (2015). State of the science: Implicit bias review. Kirwan Institute for the Study of Race and Ethnicity: The Ohio State University.

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control:

Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, *34*(10), 1332-1345.

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.

Trawalter, S., Richeson, J. A., & Shelton, J. N. (2009). Predicting behavior during interracial interactions: A stress and coping approach. *Personality and Social Psychology Review*, *13*(4), 243-268.

Webb, T. L., Sheeran, P., & Pepper, J. (2012). Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials. *British Journal of Social Psychology*, *51*(1), 13-32.

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

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.

Zogmaister, C., Arcuri, L., Castelli, L., & Smith, E. R. (2008). The impact of loyalty and equality on implicit ingroup favoritism. *Group Processes & Intergroup Relations*, *11*(4), 493-512.

Zuwerink, J. Z., Devine, P. G., Monteith, M. J., & Cook, D. A. (1996). Prejudice toward Blacks: With and without compunction? *Basic and Applied Social Psychology*, *18*(2), 131-150.

Tables

Table 1

Sensitivity: All Participants

Condition	*N*	Less Attractive	More Attractive	*t*	*p*	*d [95% CI]*
Control	700	1.04 (.64)	1.00 (.66)	1.46	.145	.06 [-.02, .13]
Accountability	609	.98 (.65)	1.01 (.69)	-1.11	.267	-.05 [-.12, .04]
Implementation Intentions	668	1.04 (.64)	1.00 (.67)	1.85	.065	.07 [-.01, .15]
Commitment	604	1.01 (.66)	1.01 (.64)	.17	.868	.01 [-.07, .09]
Awareness + Fairness	673	1.10 (.59)	1.00 (.64)	2.97	.003	.11 [.04, .19]

Criterion: All Participants

Condition	*N*	Less Attractive	More Attractive	*t*	*p*	*d [95% CI]*
Control	700	.03 (.49)	-.14 (.48)	8.75	<.001	.33 [.25, .41]
Accountability	609	.01 (.48)	-.08 (.47)	3.89	<.001	.16 [.07, .24]
Implementation Intentions	668	-.02 (.45)	-.08 (.46)	3.20	.001	.12 [.05, .20]
Commitment	604	-.02 (.49)	-.07 (.48)	2.27	.023	.09 [.01, .17]
Awareness + Fairness	673	-.06 (.45)	-.06 (.47)	-.34	.732	-.01 [-.09, .06]

Criterion: Participants Reporting Showing No Bias

Condition	*N*	Less Attractive	More Attractive	*t*	*p*	*d [95% CI]*
Control	504	-.002 (.49)	-.08 (.47)	4.19	<.001	.19 [.10, .27]
Accountability	439	-.03 (.45)	-.05 (.46)	.56	.575	.03 [-.07, .12]
Implementation Intentions	537	-.04 (.45)	-.06 (.45)	1.36	.173	.06 [-.03, .14]
Commitment	466	-.03 (.48)	-.04 (.46)	.40	.689	.02 [-.07, .11]
Awareness + Fairness	519	-.09 (.43)	-.05 (.45)	-1.97	.050	-.09 [-.17, 0]

Criterion: Participants Reporting Wanting To Show No Bias

Condition	*N*	Less Attractive	More Attractive	*t*	*p*	*d [95% CI]*
Control	587	.03 (.49)	-.13 (.48)	7.88	<.001	.33 [.24, .41]
Accountability	498	-.02 (.47)	-.06 (.47)	2.06	.041	.09 [.004, .18]
Implementation Intentions	550	-.04 (.45)	-.08 (.46)	2.12	.034	.09 [.01, .17]
Commitment	534	-.03 (.49)	-.06 (.47)	1.33	.185	.06 [-.03, .14]
Awareness + Fairness	568	-.06 (.44)	-.05 (.46)	-.61	.541	-.03 [-.11, .06]

Table 1. Sample size, means, test statistics, p-values and effect sizes for criterion and sensitivity towards less physically attractive and more physically attractive applicants within each Study 1 condition.

Table 2
Comparison in Criterion Bias Between Experimental vs. Control Condition

Condition	t	*p*	*d* [95 % CI]
Accountability	2.99	.003	.17 [.06, .27]
Implementation Intentions	4.41	<.001	.24 [.13, .34]
Commitment	4.34	<.001	.24 [.13, .35]
Awareness + Fairness	6.61	<.001	.36 [.25, .46]

Table 2. Test statistics, p values and effect sizes comparing the criterion bias in each experimental condition to Study 1's *Control* condition. Each experimental condition reduced criterion bias relative to the *Control* condition.

Table 3

Explicit and Implicit Attitudes				
Condition	*Explicit N*	*Exp. Preference*	*Implicit N*	*BIAT D*
Control	645	.72 (.96)	601	.62 (.45)
Accountability	549	.70 (.98)	521	.62 (.45)
Implementation Intentions	626	.69 (1.03)	590	.62 (.47)
Commitment	567	.63 (.93)	511	.61 (.48)
Awareness + Fairness	643	.72 (1.04)	585	.56 (.45)
Perceived and Desired Performance				
Condition	*Perceived N*	*Percent Fair*	*Desired N*	*Percent Fair*
Control	656	76.8%	650	90.3%
Accountability	571	76.9%	563	88.5%
Implementation Intentions	645	83.3%	628	87.6%
Commitment	576	80.9%	575	92.9%
Awareness + Fairness	644	80.6%	639	88.9%

Table 3. Samples sizes, means and standard deviations for measures of explicit and implicit attitudes within each experimental condition. Table also presents sample sizes and percentage of participants reporting no bias for perceived and desired performance. Exp. Preference = Explicit preference item for more vs. less physically attractive people. BIAT *D* = *D* score on Brief Implicit Attitudes Test.

Table 4
Coefficients, confidence intervals and p values for model predicting criterion bias in Study 1

Parameter	B	*95% CI*	*t*	*p*
Exp. Pref	.03	[-.01, .12]	1.55	.120
BIAT *D*	.12	[.04, .20]	2.95	.003
Accountability Condition	.33	[-.22, .88]	1.16	.245
Awareness Condition	.67	[.15, 1.23]	2.50	.013
Commitment Condition	.85	[.19, 1.50]	2.54	.011
Implementation Condition	.94	[.40, 1.48]	3.41	.001
Perc. Performance	.14	[.09, .19]	5.67	<.001
Des. Performance	.15	[.06, .23]	3.33	.001
Accountability * Explicit Preference	-.01	[-.07, .05]	-.41	.686
Awareness * Explicit Preference	.02	[-.03, .08]	.86	.390
Commitment * Explicit Preference	.01	[-.05, .07]	.29	.775
Implementation * Explicit Preference	.02	[-.03, .08]	.85	.394
Accountability * BIAT *D*	-.04	[-.16, .08]	-.63	.530
Awareness * BIAT *D*	.05	[-.07, .16]	.77	.443
Commitment * BIAT *D*	.04	[-.07, .16]	.71	.477
Implementation * BIAT *D*	-.03	[-.03, .15]	-.56	.574
Accountability * Perceived Performance	.07	[-.01, 14X]	1.87	.061
Awareness * Perceived Performance	-.14	[-.22, -.06]	-3.42	.001
Commitment * Perceived Performance	.04	[-.04, .13]	.99	.323
Implementation * Perceived Performance	-.12	[-.21, -.04]	-3.04	.002
Accountability * Desired Performance	-.15	[-.26, -.04]	-2.61	.009
Awareness * Desired Performance	-.11	[-.23, .01]	-1.87	.062
Commitment * Desired Performance	-.30	[-.44, -.16]	-4.14	<.001
Implementation * Desired Performance	-.16	[-.27, -.05]	-2.4	.006

Table 4. Study 1 output of linear regression predicting size of criterion bias by experimental condition, explicit attitudes, implicit attitudes, perceived performance and desired performance as well as interactions between condition and attitudes and measures of task performance. BIAT *D* = *D* score from Brief Implicit Association Test.

Table 5

All Eligible Participants

Condition	*N*	Less Attractive	More Attractive	*t*	*p*	*d [95% CI]*
No Awareness / No Fairness	217	-.13 (.50)	.03 (.50)	4.90	<.001	.33 [.20, .47]
No Awareness / Fairness	250	-.17 (.47)	.04 (.48)	6.00	<.001	.38 [.25, .51]
Awareness / No Fairness	234	-.14 (.51)	-.05 (.52)	2.72	.007	.18 [.05, .31]
Awareness / Fairness	239	-.11 (.49)	-.03 (.54)	2.36	.019	.15 [.03, .28]

Reported Showing No Bias

Condition	*N*	Less Attractive	More Attractive	*t*	*p*	*d [95% CI]*
No Awareness / No Fairness	150	-.10 (.50)	-.03 (.49)	2.39	.018	.19 [.03, .36]
No Awareness / Fairness	174	-.10 (.44)	.02 (.45)	3.59	<.001	.27 [.12, .42]
Awareness / No Fairness	161	-.12 (.50)	-.14 (.49)	-.58	.565	-.05 [-.20, .11]
Awareness / Fairness	166	-.08 (.47)	-.05 (.51)	.94	.351	.07 [-.08, .22]

Reported Wanting To Show No Bias

Condition	*N*	Less Attractive	More Attractive	*t*	*p*	*d [95% CI]*
No Awareness / No Fairness	182	-.13 (.48)	.001 (.50)	4.23	<.001	.31 [.16, .46]
No Awareness / Fairness	206	-.13 (.45)	.02 (.45)	4.66	<.001	.32 [.18, .46]
Awareness / No Fairness	184	-.12 (.49)	-.08 (.50)	1.44	.151	.10 [-.04, .25]
Awareness / Fairness	189	-.08 (.48)	-.02 (.54)	1.45	.150	.10 [-.04, .25]

Table 5. Sample size, means, test statistics, p-values and effect sizes for criterion towards less physically attractive and more physically attractive applicants within each Study 2 condition.

Table 6

Explicit and Implicit Attitudes				
Condition	*Explicit N*	*M (SD)*	*Implicit N*	*M (SD)*
No Awareness / No Fairness	200	.65 (.92)	184	.72 (.53)
No Awareness / Fairness	225	.81 (.91)	210	.69 (.50)
Awareness / No Fairness	217	.71 (.99)	190	.78 (.50)
Awareness / Fairness	212	.69 (1.03)	197	.62 (.49)
Perceived And Desired Performance				
Condition	*Perceived N*	*M (SD)*	*Desired N*	*M (SD)*
No Awareness / No Fairness	201	.10 (.77)	198	-.04 (.41)
No Awareness / Fairness	236	.15 (.84)	232	-.02 (.59)
Awareness / No Fairness	224	.02 (.93)	215	-.06 (.66)
Awareness / Fairness	226	.10 (.82)	221	-.06 (.60)

Table 6. Samples sizes, means and standard deviations for measures of explicit and implicit attitudes as well as perceived and desired performance within each Study 2 condition. Explicit measure = Explicit preference item for more vs. less physically attractive people. Implicit measure = *D* score on Brief Implicit Attitudes Test.

Table 7
Coefficients and confidence intervals for model predicting criterion bias in Study 2

Parameter	B	*95% CI*
Exp. Pref	.07	[.01, .14]
BIAT *D*	.07	[-.05, .19]
Awareness Condition	-.16	[-.29, -.03]
Fairness Condition	.07	[-.05, .20]
Perc. Performance	.16	[.08, .24]
Des. Performance	-.08	[-.23, .07]
Awareness * Explicit Preference	.002	[-.07, .08]
Fairness* Explicit Preference	-.05	[-.12, .03]
Awareness * BIAT *D*	.10	[-.04, .24]
Fairness * BIAT *D*	-.02	[-.16, .12]
Awareness * Perceived Performance	-.03	[-.12, .06]
Fairness * Perceived Performance	.09	[.001, .18]
Awareness * Desired Performance	.03	[-.13, .18]
Fairness * Desired Performance	.01	[-.14, .16]

Table 7. Study 2 output for linear regression predicting size of criterion bias by awareness and fairness experimental conditions, explicit attitudes, implicit attitudes, perceived performance, desired performance as well as interactions between condition with attitudes and measures of task performance. BIAT *D* = *D* score from Brief Implicit Association Test.

Table 8

All Eligible Participants

Condition	*N*	*Less Attractive*	*More Attractive*	*t*	*p*	*d [95% CI]*
No Awareness / Control	458	-.12 (.46)	.05 (.47)	8.26	<.001	.39 [.29, .48]
No Awareness / Accountable	429	-.13 (.47)	.04 (.50)	7.24	<.001	.35 [.25, .45]
No Awareness / Implement	446	-.13 (.44)	-.01 (.50)	5.49	<.001	.26 [.17, .35]
No Awareness / Objective	382	-.12 (.46)	-.004 (.45)	4.91	<.001	.25 [.15, .35]
Awareness / Control	445	-.07 (.46)	-.03 (.46)	1.53	.126	.07 [-.02, .17]
Awareness / Accountable	439	-.09 (.48)	-.02 (.48)	2.87	.004	.14 [.04, .23]
Awareness / Implement	444	-.09 (.44)	-.03 (.48)	2.77	.006	.13 [.04, .22]
Awareness / Objective	426	-.08 (.47)	-.05 (.48)	1.00	.317	.05 [-.05, .14]

Reported Showing No Bias

Condition	*N*	*Less Attractive*	*More Attractive*	*t*	*p*	*d [95% CI]*
No Awareness / Control	326	-.08 (.45)	.03 (.45)	5.12	<.001	.28 [.17, .39]
No Awareness / Accountable	301	-.10 (.48)	-.01 (.47)	4.12	<.001	.24 [.12, .35]
No Awareness / Implement	332	-.11 (.43)	-.04 (.45)	3.23	.001	.18 [.07, .29]
No Awareness / Objective	282	-.09 (.45)	-.04 (.44)	2.04	.042	.12 [.01, .24]
Awareness / Control	337	-.06 (.47)	-.05 (.45)	.18	.859	.01 [-.10, .12]
Awareness / Accountable	303	-.05 (.47)	-.05 (.48)	.06	.955	.01 [-.11, .12]
Awareness / Implement	320	-.09 (.43)	-.06 (.45)	1.52	.129	.09 [-.02, .19]
Awareness / Objective	313	-.07 (.44)	-.08 (.46)	-.23	.818	-.01 [-.12, .10]

Reported Wanting To Show No Bias

Condition	*N*	*Less Attractive*	*More Attractive*	*t*	*p*	*d [95% CI]*
No Awareness / Control	379	-.12 (.45)	.04 (.47)	7.04	<.001	.36 [.26, .47]
No Awareness / Accountable	357	-.11 (.47)	.01 (.46)	5.40	<.001	.29 [.18, .39]
No Awareness / Implement	370	-.11 (.42)	-.04 (.46)	3.65	<.001	.19 [.09, .29]
No Awareness / Objective	326	-.09 (.45)	-.02 (.44)	3.40	.001	.19 [.08, .30]
Awareness / Control	368	-.13 (.45)	.03 (.45)	.42	.673	.02 [-.08, .12]
Awareness / Accountable	352	-.17 (.46)	.04 (.46)	.92	.357	.05 [-.05, .15]
Awareness / Implement	351	-.14 (.42)	-.05 (.47)	1.78	.076	.10 [-.01, .20]
Awareness / Objective	357	-.11 (.46)	-.03 (.47)	.15	.878	.01 [-.10, .11]

Table 8. Sample size, means, test statistics, p-values and effect sizes for criterion towards less physically attractive and more physically attractive applicants within each Study 3 condition.

Table 9

Explicit and Implicit Attitudes				
Condition	*Explicit N*	*M (SD)*	*Implicit N*	*M (SD)*
No Awareness / Control	421	.72 (.97)	399	.69 (.50)
No Awareness / Accountable	381	.67 (1.02)	365	.71 (.50)
No Awareness / Implement	416	.74 (.97)	397	.71 (.50)
No Awareness / Objective	365	.67 (.94)	346	.70 (.53)
Awareness / Control	411	.72 (1.02)	388	.69 (.52)
Awareness / Accountable	397	.74 (.99)	364	.67 (.52)
Awareness / Implement	397	.77 (1.01)	381	.69 (.49)
Awareness / Objective	396	.70 (.89)	372	.71 (.48)
Perceived And Desired Performance				
Condition	*Perceived N*	*M (SD)*	*Desired N*	*M (SD)*
No Awareness / Control	430	.17 (.73)	422	.02 (.55)
No Awareness / Accountable	394	.05 (.75)	391	-.03 (.50)
No Awareness / Implement	424	.04 (.72)	418	-.02 (.50)
No Awareness / Objective	367	.03 (.76)	365	0 (.54)
Awareness / Control	418	.02 (.71)	408	-.02 (.49)
Awareness / Accountable	399	.03 (.71)	399	-.01 (.49)
Awareness / Implement	416	.04 (.78)	406	-.05 (.58)
Awareness / Objective	405	-.04 (.65)	404	-.04 (.53)

Table 9. Samples sizes, means and standard deviations for measures of explicit and implicit attitudes as well as perceived and desired performance within each Study 3 condition. Explicit measure = Explicit preference item for more vs. less physically attractive people. Implicit measure = *D* score on Brief Implicit Attitudes Test.

Table 10
Coefficients and confidence intervals for model predicting criterion bias in Study 3

Parameter	B	*95% CI*
Exp. Pref	.02	[-.02, .05]
BIAT *D*	-.02	[-.09, .05]
Perc. Performance	.14	[.09, .20]
Des. Performance	.003	[-.07, .08]
Awareness Condition	-.17	[-.22, -.11]
Accountability Condition	-.05	[-.13, .03]
Implementation Condition	-.01	[-.09, .07]
Objectivity Condition	-.12	[-.20, -.04]
Awareness * Explicit Preference	.03	[.01, .07]
Accountability* Explicit Preference	.02	[-.03, .07]
Implementation * Explicit Preference	-.002	[-.05, .05]
Objectivity* Explicit Preference	.04	[-.01, .09]
Awareness * BIAT *D*	.09	[.03, .15]
Accountability* BIAT *D*	.04	[-.05, .13]
Implementation * BIAT *D*	.005	[-.08, .09]
Objectivity* BIAT *D*	.09	[.01, .18]
Awareness * Perceived Performance	.01	[-.05, .06]
Accountability* Perceived Performance	-.003	[-.07, .07]
Implementation * Perceived Performance	-.02	[-.09, .05]
Objectivity* Perceived Performance	.04	[-.03, .11]
Awareness * Desired Performance	.01	[-.06, .08]
Accountability* Desired Performance	.02	[-.08, .12]
Implementation * Desired Performance	.10	[.01, .19]
Objectivity* Desired Performance	-.01	[-.10, .08]

Table 10. Study 3 output for linear regression predicting size of criterion bias by awareness and manipulation experimental conditions, explicit attitudes, implicit attitudes, perceived performance, desired performance as well as interactions between condition with attitudes and measures of task performance. BIAT *D* = *D* score from Brief Implicit Association Test.

Table 11

2 x 2 ANOVAs for both Control and Awareness conditions in Study 4a and Study 4b.

Term	df	*F*	*p*	η^2_p
Study 4a				
Control Condition				
Attractiveness	(1, 424)	80.37	<.001	.159
School	(1, 424)	9.25	.002	.021
Attractiveness * School	(1, 424)	5.48	.020	.013
Awareness Condition				
Attractiveness	(1, 488)	48.96	<.001	.091
School	(1, 488)	0.87	.421	.002
Attractiveness * School	(1, 488)	0.10	.809	<.001
Study 4b				
Control Condition				
Attractiveness	(1, 466)	44.65	<.001	.087
Political Group	(1, 466)	55.78	<.001	.107
Attractiveness * Political Group	(1, 466)	0.24	.622	.001
Awareness Condition				
Attractiveness	(1, 432)	61.70	<.001	.125
Political Group	(1, 432)	14.18	<.001	.032
Attractiveness * Political Group	(1, 432)	0.78	.377	.002

Table 11. Output for 2 (Attractiveness) by 2 (Group) ANOVAs on response criterion within each experimental condition for Studies 4a and 4b.

Table 12

Descriptive and test statistics for attitude and performance measures in Studies 4a-4b.

Measure	Control *N*	Control *M* (SD)	Aware *N*	Aware *M* (SD)	*t*	*p*	*d* [95% CI]
Study 4a							
Attractiveness Measures							
Exp. Pref	425	1.01 (.82)	489	1.07 (.85)	-1.00	.318	-.07 [-.20, .06]
IAT *D*	422	.73 (.34)	485	.74 (.33)	-0.43	.665	-.03 [-.16, .10]
Perc. Perf	425	.39 (.69)	489	.31 (.60)	1.68	.094	.11 [-.02, .24]
Des. Perf	425	.07 (.47)	489	.08 (.43)	-0.44	.658	-.03 [-.16, .10]
School Measures							
Exp. Pref	425	.72 (.91)	489	.66 (.88)	1.03	.301	.07 [-.06, .20]
IAT *D*	422	.41 (.32)	485	.39 (.29)	1.12	.265	.07 [-.06,.20]
Perc. Perf	425	.19 (.49)	489	.10 (.40)	2.98	.003	.20 [.07, .33]
Des. Perf	425	.17 (.54)	489	.06 (.36)	3.56	<.001	.24 [.11, .37]
Study 4b							
Attractiveness Measures							
Exp. Pref	449	.44 (1.03)	423	.38 (.95)	0.94	.349	.06 [-.07,.20]
Perc. Perf	454	.05 (.83)	425	.09 (.77)	-0.68	.498	-.05 [-.18,.09]
Des. Perf	457	.04 (.76)	424	.09 (.61)	-0.98	.325	-.07 [-.20,.07]
Politics Measures							
Exp. Pref	462	1.22 (1.28)	430	1.03 (1.16)	2.32	.021	.16 [.02,.29]
Perc. Perf	460	.32 (.87)	427	.20 (.73)	2.22	.026	.15 [.02,.28]
Des. Perf	461	.32 (.81)	427	.24 (.74)	1.54	.125	.10 [-.03,.23]

Table 12. Samples sizes, means and standard deviations for attitude and performance measures within each experimental condition. Exp. Pref = Explicit preference item. IAT *D* = *D* score from Implicit Association Test. Perc. Perf = Perceived performance item. Des. Perf = Desired performance item.

Table 13
Coefficients and confidence intervals for models predicting criterion bias in Studies 4a and 4b

Study 4a		
Attractiveness Bias		
Parameter	B	95% CI
Explicit Preference	.03	[.001, .07]
IAT *D*	.07	[-.01, .15]
Perceived Performance	.14	[.10, .19]
Desired Performance	.05	[-.01, .11]
School Bias		
Explicit Preference	.02	[-.01, .05]
IAT *D*	.05	[-.03, .13]
Perceived Performance	.08	[.03, .14]
Desired Performance	.06	[.004, .12]
Study 4b		
Attractiveness Bias		
Explicit Preference	.05	[.01, .08]
Perceived Performance	-.01	[-.06, .04]
Desired Performance	-.01	[-.07, .05]
Political Group Bias		
Explicit Preference	.05	[.01, .09]
Perceived Performance	.26	[.18, .34]
Desired Performance	.19	[.11, .27]

Table 13. Study 4a and 4b output for linear regression predicting size of criterion bias by explicit attitudes, implicit attitudes, perceived performance, and desired performance. IAT *D* = *D* score from Implicit Association Test.

Table 14
Coefficients and confidence intervals for models predicting criterion bias in Studies 4a and 4b

Parameter	B	95% CI
Study 4a		
Attractiveness Bias		
Condition	-.10	[-.24, .04]
Explicit Preference	.05	[-.01, .10]
IAT *D*	-.01	[-.12, .11]
Perceived Performance	.19	[.13, .25]
Desired Performance	.08	[-.01, .16]
Explicit Preference* Condition	-.02	[-.09, .05]
IAT *D* * Condition	.16	[.001, .32]
Perceived Performance * Condition	-.10	[-.19, -.01]
Desired Performance * Condition	-.05	[-.17, .07]
School Bias		
Condition	-.03	[-.12, .05]
Explicit Preference	.03	[-.02, .07]
IAT *D*	.01	[-.11, .12]
Perceived Performance	.11	[.03, .19]
Desired Performance	.08	[.01, .15]
Explicit Preference* Condition	-.02	[-.08, .04]
IAT *D* * Condition	.09	[-.07, .25]
Perceived Performance * Condition	-.07	[-.19, .04]
Desired Performance * Condition	-.09	[-.21, .03]
Study 4b		
Attractiveness Bias		
Condition	.01	[-.06, .08]
Explicit Preference	.02	[-.02, .07]
Perceived Performance	.02	[-.02, .07]
Desired Performance	.01	[-.07, .08]
Explicit Preference* Condition	.05	[-.01, .12]
Perceived Performance * Condition	-.05	[-.15, .06]
Desired Performance * Condition	-.04	[-.17, .09]
Political Group Bias		
Condition	.04	[-.08, .16]
Explicit Preference	.09	[.04, .14]
Perceived Performance	.17	[.07, .27]

Desired Performance	.39	[.29, .49]
Explicit Preference* Condition	-.09	[-.17, -.01]
Perceived Performance * Condition	.17	[.02, .32]
Desired Performance * Condition	-.43	[-.59, -.28]

Table 14. Study 4a and 4b output for linear regression predicting size of criterion bias by experimental condition, explicit attitudes, implicit attitudes, perceived performance, desired performance as well as interactions between condition and attitudes and measures of task performance. IAT *D* = *D* score from Implicit Association Test.

Appendix A

Less Qualified Applications

Science GPA	Humanities GPA	Recommendation Letter	Interview Score
3.3	3.4	Good	82.5
3.2	3.3	Excellent	62.5
3.5	3.6	Good	72.5
3.7	3.5	Good	70
3.1	3.4	Excellent	62.5
3.2	3.7	Good	77.5
3.8	3.3	Good	72.5
3.3	3.2	Good	87.5
3.0	3.3	Excellent	67.5
3.6	3.1	Good	82.5
3.7	3.2	Good	77.5
3.3	3.4	Excellent	57.5
3.5	3.4	Good	77.5
3.8	3.1	Good	77.5
3.1	3.7	Good	80
3.5	3.7	Good	70
3.6	3.3	Excellent	52.5
3.2	3.4	Good	85
3.6	3.7	Good	67.5
3.7	3.4	Good	72.5
3.1	3.6	Good	82.5
3.5	3.0	Excellent	62.5
3.2	3.1	Good	92.5
3.9	3.2	Good	72.5
3.0	3.1	Excellent	72.5
3.5	3.9	Good	65
3.4	3.4	Good	80
3.2	3.1	Excellent	67.5
3.8	3.4	Good	70
3.1	3.5	Excellent	60
3.8	3.0	Good	80
3.3	3.1	Excellent	65

More Qualified Applications

Science GPA	Humanities GPA	Recommendation Letter	Interview Score
3.8	3.3	Good	97.5
3.2	3.4	Excellent	85
3.9	3.7	Good	85
3.2	3.7	Excellent	77.5
3.5	3.5	Excellent	75
2.9	3.4	Excellent	92.5
3.8	3.0	Excellent	80
3.6	3.4	Excellent	75
3.7	3.9	Good	85
3.4	3.6	Excellent	75
3.1	3.2	Excellent	92.5
3.6	3.7	Good	92.5
3.4	3.0	Excellent	90
3.4	3.5	Excellent	77.5
3.3	3.2	Excellent	87.5
3.5	3.4	Excellent	77.5
3.7	3.8	Good	87.5
3.8	3.6	Good	90
3.5	3.3	Excellent	80
3.2	3.9	Excellent	72.5
3.8	3.4	Excellent	70
3.1	3.4	Excellent	87.5
3.3	3.7	Excellent	75
3.8	3.7	Good	87.5
3.9	3.8	Good	82.5
3.6	3.7	Excellent	67.5
3.3	3.6	Excellent	77.5
3.8	3.4	Good	95
3.5	3.7	Excellent	70
3.7	3.6	Excellent	67.5
3.8	3.8	Good	85
3.2	3.2	Excellent	90

Appendix B

Procedural details for the four-block Brief Implicit Association Test used in Studies 1-3

Each BIAT block had 20 trials, with the first four trials in each block being practice trials and were not analyzed. In each block, participants pressed the "i" key for Good words (Love, Pleasant, Great, Wonderful) and either more attractive or less attractive faces, and the "e" for "items that do not belong to these categories", which consisted of Bad words (Hate, Unpleasant, Awful, Terrible) and images of whatever attractiveness category was not paired with good words.

Blocks 1 and 3 always had the same pairings, as did blocks 2 and 4. Participants were randomly assigned to an order that paired Good words with more attractive faces first or an order that paired Good words with less attractive faces first.

Appendix C

Text for Awareness manipulation in Study 4a:

The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.

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 university and tougher on candidates from another university.

Can you be fair toward all applicants and not be biased by their university? 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 candidates from your own university over candidates from another university. When you are done, please type this strategy in the box below.

[Text seen immediately before testing]

Please remember to be fair and avoid favoring candidates from your own university over candidates from another university.

Text for Awareness manipulation in Study 4b:

The goal of this study is to learn about decision-making. You will determine whether or not each applicant should be accepted into an honor society based on his or her science GPA, humanities GPA, recommendation letter, and interview score.

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.

[Text seen immediately before testing]

Remember to avoid favoring candidates from your own political party over candidates from other political parties.

Appendix D

Implicit Association Procedure for Study 4a:

Implicit school and attractiveness attitudes were assessed using a 7-block Implicit Association Test (IAT; Greenwald et al., 1998), measuring association strengths between the categories of either 1) University of Virginia (UVA) and University of North Carolina (UNC) and the White American and Black American and the attributes Good and Bad or 2) More Attractive People and Less Attractive People and the attributes Good and Bad. For each school, stimuli were four logos. For more and less attractive people, each category contained four faces (two male, two female) that were pre-tested as being the most and least physically attractive of the images used in the JBT.

Within each IAT, participants were randomly assigned to use left or right keys for each category or attribute as well as being randomly assigned to complete either congruent or incongruent blocks first.

In the first block (practice, 20 trials), participants categorize only images from two categories using the "e" and "i" keys. In the second block (practice, 20 trials), participants categorize only words from the two attributes: Good words (pleasant, great, wonderful, excellent) and Bad words (hate, unpleasant, awful, terrible). In the third block (test, 20 trials) and fourth block (test, 40 trials) participants must categorizes both images from one category and words from one attribute jointly using the same key (e.g., images of More Attractive People and Good words with one key, images of Less Attractive People and Bad words with the other key).

In fifth block (practice, 20 trials), participants categorize only images from the two categories, using the opposite keys from those assigned in the first block. Finally, in the sixth (test, 20 trials) and seventh (test, 40 trials) blocks, participants categorize both images and words from

one category and attribute using the same key, now completing the opposite pairing of that in the third and fourth blocks (e.g., images of Less Attractive People and Good words with one key, images of More Attractive People and Bad words with the other key).

The IATs were scored according following the guidelines of Greenwald, Nosek, and Banaji (2003) such that more positive values indicated a stronger implicit association between 1) UVA with Good and UNC with Bad or 2) More Attractive people with Good and Less Attractive People with Bad). IAT scores were retained if fewer than 10% of the response trials had a latency less than 300 milliseconds, as recommended in Nosek et al., 2007.