Journal of Experimental Social Psychology 76 (2018) 337–355

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

The Judgment Bias Task: A flexible method for assessing individual differences in social judgment biases

Jordan R. Axt[a,*], Helen Nguyen[a], Brian A. Nosek[a,b]

[a] University of Virginia, United States
[b] Center for Open Science, United States

ARTICLE INFO

Handling editor: Elizabeth Page-Gould

Keywords:
Bias
Discrimination
Judgment
Prejudice
Favoritism

ABSTRACT

Many areas of social psychological research investigate how social information may bias judgment. However, most measures of social judgment biases are (1) low in reliability because they use a single response, (2) not indicative of individual differences in bias because they use between-subjects designs, (3) inflexible because they are designed for a particular domain, and (4) ambiguous about magnitude of bias because there is no objectively correct answer. We developed a measure of social judgment bias, the Judgment Bias Task, in which participants judge profiles varying in quality for a certain outcome based on objective criteria. The presence of ostensibly irrelevant social information provides opportunity to assess the extent to which social biases undermine the use of objective criteria in judgment. The JBT facilitates measurement of social judgment biases by (1) using multiple responses, (2) indicating individual differences by using within-subject designs, (3) being adaptable for assessing a variety of judgments, (4) identifying an objective magnitude of bias, and (5) taking 6 min to complete on average. In nine pre-registered studies ($N > 9000$) we use the JBT to reveal two prominent social judgment biases: favoritism towards more physically attractive people and towards members of one's ingroup. We observe that the JBT can reveal social biases, and that these sometimes occur even when the participant did not intend or believe they showed biased judgment. A flexible, objective, efficient assessment of social judgment biases will accelerate theoretical and empirical progress.

1. Introduction

Social bias – intended or unintended favoritism in evaluation, judgment, or behavior for one social group over another – is pervasive. Sometimes people are aware of their biases and embrace them as guides for behavior. For example, the first author only watches Duke basketball games with people willing to cheer for Duke, disqualifying the second and third authors. Other times, biases differ from conscious values, and can cause actions to deviate from intended behaviors. Discrimination in hiring (Ameri et al., 2015), academic (Milkman, Akinola, & Chugh, 2012), and economic (Doleac & Stein, 2013; Edelman, Luca, & Svirsky, 2017) contexts may occur without conscious intention to discriminate, or awareness of doing so (Bertrand, Chugh, & Mullainathan, 2005; Bertrand & Duflo, 2016; Rooth, 2010).

The social consequences of biases, combined with the possibility that some occur outside of intention or awareness, have made them a popular topic of research. At the same time, there are pervasive methodological limitations for conducting controlled experimental research on judgment biases including low reliability, lack of insight on individual differences in degree of bias, lack of an objective standard indicating no bias, and idiosyncratic paradigms that cannot be adapted for multiple uses.

Low reliability. Most bias investigations rely on a single judgment or behavior as the dependent variable. In 2015, there were 68 studies testing a judgment or behavioral preference for one social group over another published in four social psychology journals: *Journal of Personality and Social Psychology*, *Personality and Social Psychology Bulletin*, *Journal of Experimental Social Psychology*, *and Psychological Science*.[1] Of them, 47 (68%) relied on only a single judgment or behavior for bias assessment, and 57 (83%) relied on five or fewer. Examples of single-shot outcomes included allocating resources (Binning, Brick, Cohen, & Sherman, 2015) or providing hypothetical prison sentences (Cheung & Heine, 2015). Single response assessments, particularly of social judgments or behaviors that are influenced by a variety of factors, are often unreliable and weaken power to detect biases. Underpowered research increases the rate of Type 1 and Type 2 errors (Button

* Corresponding author at: University of Virginia, Box 400400, Charlottesville, VA 22904-4400, United States.
E-mail address: jra3ee@virginia.edu (J.R. Axt).

[1] We included all studies that used actual behavior or behavioral intentions between existing social groups. See https://osf.io/u2mbx/ for list of all included studies and outcome measures.

https://doi.org/10.1016/j.jesp.2018.02.011
Received 9 August 2017; Received in revised form 15 February 2018; Accepted 26 February 2018
Available online 08 March 2018
0022-1031/

et al., 2013) and contributes to weakening reproducibility of research (Asendorpf et al., 2013; Funder et al., 2014).

Measuring individual differences. Many existing bias paradigms are unable to distinguish the relative strength of biased behavior between participants. Partly this is a function of lower reliability based on single responses. Another contributor is reliance on between-subjects designs. For example, in Norton, Vandello, and Darley (2004), participants chose between two fictional college applicants. Candidates had different strengths, with one applicant being Black and the other White, and race randomly assigned to strengths between subjects. Black applicants were favored regardless of condition, indicating racial bias in the aggregate. These studies were not focused on finding individual differences in social judgment bias, but there may be added benefits to developing within-subjects measures to estimate the social bias for each participant. Such a design enables assessment of group-level differences (e.g., the impact of an intervention on reducing levels of racial bias in judgment) and individual differences (e.g., the relation between racial attitudes and racial bias in judgment).

Objective standard. Many measures of bias have no objectively correct answer, meaning bias can only be understood in relative terms between participants or conditions (e.g., Blommaert, van Tubergen, & Coenders, 2012). For example, Haddock, Zanna, and Esses (1993) used a hypothetical budget paradigm to study attitudes towards gay people. Participants needed to cut funding for several organizations, one of which was the university's gay and lesbian organization. More prejudiced participants proposed harsher reductions in funding towards the gay and lesbian organization. However, there is no objective standard for what level of funding indicates lack of bias. As a consequence, there is no way to identify who is biased and to what extent they are biased.

It is often of practical, legal, and theoretical interest to know if social judgments conform to an objective standard. If measures can only represent biased behavior in relative terms, then it is not possible to investigate or conclude when a judgment or behavior is unbiased.

Adaptability for multiple uses. Implicit measures like the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2007) are used frequently, in part, because they can be adapted to a variety of domains. To measure new content, researchers retain the established procedural parameters and change just the task stimuli following established best practices (Lane, Banaji, Nosek, & Greenwald, 2007). For many social judgment bias measures, the procedure and content are not easily separated, making it difficult to adapt the method for other uses. For example, measures investigating social bias through employment resumes cannot be easily adapted to other forms of social bias. Moreover, measures like the IAT are reliable and efficient to administer by collecting multiple responses quickly, which maximizes applicability across research contexts.

Given limitations of existing measures, we sought to develop a measure of social judgment bias that (1) maximized effective reliability, (2) is sensitive to measuring well-known biases, (3) identified individual differences in bias, (4) can identify magnitude of bias compared to an objective standard, (5) is efficient to administer, and (6) is flexible for a variety of uses.

1.1. The Judgment Bias Task

Prior studies on intergroup bias used methods that share some of the intended strengths of the Judgment Bias Task (JBT). For example, some studies asked participants to predict individuals' future behavior based on profiles that included both diagnostic information and irrelevant social information (e.g., gender; Beckett & Park, 1995; Locksley, Hepburn, & Ortiz, 1982). Equating the diagnostic information across social categories enabled assessment of the impact of the social information in forming predictions. Likewise, conjoint analysis reveals social bias by asking participants to choose between multiple pairs of targets who vary on levels of both task-relevant information and task-irrelevant social information (e.g., perceived weight; Caruso, Rahnev, & Banaji, 2009). By equating targets on task-relevant information across social groups, conjoint analysis can reveal the extent to which social information influences choices. Finally, Situational Judgment Tests (SJTs), common in personnel psychology (e.g., Cabrera & Nguyen, 2001), present participants with hypothetical and ambiguous scenarios and ask them to rank potential responses. Researchers can design SJTs to measure social judgment biases often not aware to participants.

The JBT builds on some of the features of these paradigms to assess social judgment biases. In the JBT, participants evaluate a series of profiles for a particular outcome, such as membership in an honor society or selection of team members. Each profile has multiple quantified criteria that are relevant for decision-making and one or more that are ostensibly irrelevant. Participants are instructed to weigh the relevant criteria equally in their judgment. The profiles are constructed so that some are systematically better than the others, but the difference is somewhat difficult to detect. Participants are assessed on their sensitivity to distinguishing between the better and worse profiles, and whether they have a bias to be more lenient or stringent to candidates with different irrelevant criteria.

One example of a JBT involves instructing participants to accept approximately half of the applicants to a hypothetical honor society. Each applicant profile has four pieces of relevant information: Science GPA, Humanities GPA, recommendation letter strength, and interview score. Simultaneously, ostensibly irrelevant gender information is communicated with a face accompanying the profile. Unobtrusively, a random half of the male and female profiles are made somewhat more qualified than the others. Participants then evaluate the individual profiles sequentially to make accept-reject decisions. Each participant's performance produces scores for their ability to distinguish more from less qualified applicants, and whether judgments were more lenient or strict compared to the objective standard, both overall and separately for each gender.

Unlike past work using related methods investigating intergroup bias (Beckett & Park, 1995; Cabrera & Nguyen, 2001; Locksley et al., 1982), the JBT is analyzed using Signal Detection Theory (SDT). Decisions made during the task can be assessed based on *sensitivity* (d′) and *criterion* (c). Sensitivity measures the extent to which a participant distinguishes more from less qualified profiles. Participants with high sensitivity are better at accepting the more qualified and rejecting the less qualified profiles than those with low sensitivity. A score of zero indicates no ability to distinguish more from less qualified profiles.

Criterion measures the extent to which a participant is lenient or strict in evaluation. Lower criterion values indicate being more lenient, and higher criterion values indicate being more strict. A score of zero indicates equal likelihood of correctly accepting more qualified profiles and correctly rejecting less qualified profiles. By computing separate sensitivity and criterion estimates for each of the social groups in the task, the JBT measures whether participants are better able at discriminating between more and less qualified profiles and whether the criterion for acceptance differs between social groups. SDT has been used productively in implicit measures of bias such as the Go/No-Go Association Task (Nosek & Banaji, 2001) and “shooter bias” tasks (Correll, Park, Judd, & Wittenbrink, 2002).

Participants may show socially biased judgment on the JBT for a variety of reasons. For example, in a JBT assessing gender biases in academic honor society admissions, some participants may have a lower acceptance criterion for male than female applicants because they believe males are more academically gifted than females, or because they simply prefer males to females. In these cases, bias on the task is intentional. Alternatively, some participants may have a lower acceptance criterion for male than female applicants even if they wanted to treat applicants from both genders equally and believe they did so. In these cases, participants' judgments may be shaped by processes operating outside of conscious awareness or intention, such as prominent, culturally-based associations between gender and

intelligence.

Both intended and unintended social biases in judgment can contribute to disparities between groups (e.g., Bertrand et al., 2005; Forscher, Cox, Graetz, & Devine, 2015), and performance on the JBT alone cannot distinguish the extent to which biases were intended versus unintended. However, since the potential to reveal such unintended biases may be a useful application of the JBT, we measured participants' perceived and desired performance on the task, as well as their implicit and explicit attitudes towards each group included in the JBT. These additional measures allow for an investigation into the extent to which biases on the JBT were related to attitudes, and if bias on the task emerged among participants who wanted to behave in an unbiased measure and believed they had done so.

Here, we investigated bias in criterion and sensitivity towards more vs. less physically attractive people (Studies 1a-1d & Study 5) and ingroup vs. outgroup members (Studies 2–4).

2. Study 1a

There is a pervasive bias favoring physical attractiveness. In one meta-analysis, physically attractive people were judged to have more positive personality traits and life outcomes than unattractive people (average Cohen's $d = 0.61$; Feingold, 1992), despite the fact that that there is no link between attractiveness and dominance, general mental health, or intelligence. Moreover, physically attractive people tend to receive more favorable treatment in hiring and admissions (Beehr & Gilmore, 1982; Cash & Kilcullen, 1985; Hosoda, Stone-Romero, & Coats, 2003; Johnson, Podratz, Dipboye, & Gibbons, 2010). In Study 1a, we tested whether the JBT could detect favoritism for physical attractiveness in selection for an honor society.

2.1. Methods

2.1.1. Participants

We sought to collect 200 participants to have > 80% power at detecting a small within-subjects effect size of Cohen's $d = 0.2$. All studies used G*Power 3.1 to determine power and sample size. Due to overscheduling, our sample was slightly larger: 206 University of Virginia (UVA) undergraduates ($M_{age} = 18.62$, $SD = 1.47$; 63.1% White, 71.8% women) completed the study for partial course credit. All studies were approved by UVA's Institutional Review Board, and participants provided consent at the beginning of each study.

2.1.2. Design

The study used a 2 (applicant gender: male or female) × 2 (physical attractiveness: more or less) within-subjects design.

2.1.3. Procedure

Participants were run in groups of one to four at individual computer carrels. Participants completed measures in the following order: academic decision making task (JBT), a survey about task performance, explicit and implicit attitude measures in a randomized order, and demographics. See https://osf.io/tn3mz/ for the study's pre-registration and https://osf.io/u2mbx/ for materials and data from all studies.[2] For all studies, we report all measures, manipulations, and exclusion criteria.

2.1.3.1. Academic decision-making task. Participants were instructed that they would evaluate applicants to an academic honor society. Their task was to accept the most qualified and reject the least qualified applicants, and that they should accept approximately half of the applicants.

In the viewing phase, participants passively viewed 64 applicants for 1 s each in a randomized order. This provided insight on the range of qualifications among the applicants. Next, in the selection phase, participants viewed each applicant one at a time in a random order and made accept or reject decisions on each. Participants clicked on a green "Accept" square or a red "Reject" square to make their decisions. There was no time limit.

Each application had a picture of the applicant and four pieces of information: science GPA (Scale of 1–4), humanities GPA (1–4), recommendation letters (poor, fair, good, or excellent), and interview score (1–100). Participants were instructed to weigh each piece of information during evaluation.

We varied these qualifications to create 64 unique applicants; 32 were made to be *more qualified* and 32 to be *less qualified*. To determine qualification, the four pieces of applicant information were converted to a scale with a maximum score of four.[3] 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 each applicant, the four scores were summed to determine their qualifications. *Less qualified* applicants added to 13 and *more qualified* applicants added to 14. For example, a sample *less qualified* applicant had a science GPA of 3.5, humanities GPA of 3.7, a recommendation letter rating of Good, and an interview score of 70. Using the transformation explained above, this information summed to 13 (3.5 + 3.7 + (Good = 3) + (70 / 25 = 2.8) = 13). See the online supplement for the qualifications of all applicants.

We collected a large sample of potential applicant photos online and had six research assistants nominate those that were most and least physically attractive. We used the 64 most frequently nominated photos for the final set of 32 more attractive (16 male, 16 female) and 32 less attractive (16 males, 16 female) photos. These 64 photos were pretested in a pilot study of undergraduates ($N = 63$, 39 female) who rated each photo on a five-point scale of physical attractiveness (1 = Not at all, 5 = Extremely). Using a within-subjects comparison, the more attractive photos ($M = 3.45$, $SD = 0.62$) were rated as more attractive than the less attractive photos ($M = 1.50$, $SD = 0.57$), $t(62) = 20.95$, $p < .001$, $d = 2.64$, 95% C.I. [2.10, 3.16]. Furthermore, every image in the attractive set was rated as more attractive on average than every image in the unattractive set.

During the task, photos associated with each application were randomly paired such that profiles from each level of qualification were matched with 16 (8 male, 8 female) more or less attractive faces.

2.1.3.2. Perception of performance. Participants answered two items about task performance. Participants first reported perceived performance, using a seven-point scale ranging from "I was extremely easier on physically unattractive applicants and extremely tougher on physically attractive applicants" (−3) to "I was extremely easier on physically attractive applicants and extremely tougher on physically unattractive applicants" (+3), with a neutral response of "I treated both physically unattractive and physically attractive applicants equally" (0). In all studies, survey items of this format had labels for every scale point. Participants next reported desired performance, using a similar seven-point scale ranging from "I wanted to be extremely easier on physically unattractive applicants and extremely tougher on physically attractive applicants" (−3) to "I wanted to be extremely easier on physically attractive applicants and extremely tougher on physically unattractive applicants" (+3), and a neutral midpoint of "I wanted to treat both physically unattractive and physically attractive

[2] For privacy concerns, we did not post images for Studies 1a–1d & 5, but contact the first author to use these materials for research purposes.

[3] Since interview scores could only have whole-number values, the four qualification scores could not have the same standard deviation across applicants and produce 64 unique combinations. Profiles were made to have similar standard deviations between science ($SD = 0.27$) and humanities GPA ($SD = 0.25$), as well as between recommendation letter ($SD = 0.50$) and interview scores ($SD = 0.40$).

applicants equally" (0).

2.1.3.3. Explicit preferences. Participants reported preference for physically attractive and unattractive people using a seven-point scale ranging from "I strongly prefer physically unattractive to physically attractive people" (−3) to "I strongly prefer physically attractive to physically unattractive people" (+3), and a neutral response of "I like physically unattractive and physically attractive people equally" (0).

2.1.3.4. Implicit preferences. Participants completed a seven-block IAT measuring strength of association between the concepts "Pleasant" and "Unpleasant" and the categories "More attractive people," and "Less attractive people". The stimuli were the two highest and lowest-rated male and female faces for each gender from the pilot study. IAT responses were scored by the *D* algorithm (Greenwald, Nosek, & Banaji, 2003), such that more positive scores reflected a stronger association between more attractive people and pleasant, and less attractive people and unpleasant. The procedure followed the recommended design and exclusion criteria from Nosek, Greenwald, and Banaji (2005). Procedural details are available in the online supplement.

2.1.3.5. Demographics. Participants completed a seven-item demographics questionnaire. We only analyzed the gender, age and race items.

2.2. Results

In Study 1a, we first examined differences in criterion among all eligible participants, and then separately among those participants who reported a desire of being unbiased, a perception of having been unbiased, and no explicit preferences between more and less physically attractive people. We then analyzed how biases in criterion related to explicit attitudes, implicit attitudes, perceived performance, and desired performance. For all studies, we report all confirmatory tests following our pre-registered analysis plan. Any deviations from that plan, or other exploratory analyses, are identified explicitly in our Results sections. Further, exploratory analyses are reported without *p*-values to emphasize the loss of diagnosticity of statistical inferences (Nosek, Ebersole, DeHaven, & Mellor, in press).

For all studies, participants were excluded from analysis if they accepted < 20% or > 80% of the applicants on the JBT, indicating a failure to follow instructions to accept approximately half. Participants were also excluded if they accepted or rejected every more attractive or less attractive applicant, indicating possible deliberately exaggerated bias. Two participants were excluded in Study 1 for these criteria. No participants had > 10% of IAT trial responses < 300 ms that would have led to excluding the IAT data (Nosek et al., 2005).

Accuracy is defined as selecting more qualified candidates and rejecting less qualified candidates. Accuracy on the task was 70.4% (SD = 7.3), above chance but not so high that identification was too easy. The average acceptance rate was close to the recommended 50% (M = 52.7%, SD = 10.2). Participants required 5.57 min on average (SD = 1.64) to complete the entire task, including reading instructions and previewing applicants.

2.2.1. Bias in response criterion

Of primary interest was the difference in response criterion for more versus less attractive applicants. More attractive applicants (M = −0.20, SD = 0.41) received a lower criterion than less attractive applicants (M = 0.04, SD = 0.38), $t(203) = 8.19$, $p < .001$, $d = 0.57$, 95% C.I. [0.42, 0.72].[4] See Fig. 1 for density plots of criterion values for Studies 1a-1d.

This criterion bias illustrates favoritism towards more physically attractive people regardless of qualifications. When more qualified, applicants were more likely to be correctly accepted if more (76.3% accuracy) than less attractive (69.9% accuracy). When less qualified, applicants were more likely to be incorrectly accepted if more (36.6% errors) than less attractive (28.0% errors). There were no reliable interactions between criterion and participant or applicant gender (see online supplement). In an exploratory analysis, sensitivity (d') was similar between more attractive (M = 1.18, SD = 0.54) and less attractive (M = 1.25, SD = 0.67) applicants, $d = 0.09$, 95% CI [−0.05, 0.23].

To measure internal reliability, the 16 trials that participants completed for each combination of qualification and attractiveness were placed into alternating sets (e.g., the first more attractive, qualified applicant judged was in the first set, the second more attractive, qualified applicant judged was in the second set, etc.) and separate criterion, as well as a criterion bias difference score, were computed for each set. We then computed a split-half reliability based on these data for both the individual criterion for each social group and the criterion difference score.

The internal reliability of the criterion measure was comparable for more (α = 0.61) and less attractive (α = 0.62) applicants. The reliability of the criterion difference score, used in the individual difference analyses below, was α = 0.33. Across studies, difference score reliabilities were lower than those of the component criterion scores, as is observed across many contexts and may underestimate effective reliability of difference scores (Williams & Zimmerman, 1996). We address the issue of reliability in the General Discussion.

2.2.2. Predicting bias in criterion

IAT *D* scores (M = 0.78, SD = 0.37, d = 2.11) and the explicit preference item (M = 1.33, SD = 0.82, d = 1.62) indicated preference towards more over less attractive people.

We computed a criterion difference score (less attractive criterion – more attractive criterion), such that higher values meant lower criterion for more versus less attractive applicants. This criterion difference score was positively correlated with IAT *D* scores ($r(204) = 0.15$, $p = .028$, 95% C.I. [0.02, 0.28]), perceptions of performance ($r(204) = 0.30$, $p < .001$, 95% C.I. [0.17, 0.42]), and desired performance ($r(204) = 0.29$, $p < .001$, 95% C.I. [0.16, 0.41]). These positive correlations indicate that participants who had more positive implicit attitudes towards more attractive people, a greater desire to favor more attractive people, and a greater perception of having favored more attractive people were more likely to have a more relaxed criterion for more attractive relative to less attractive applicants. Criterion bias was not reliably related to explicit attitudes ($r(204) = 0.09$, $p = .216$, 95% C.I. [−0.05, 0.22]).

A simultaneous linear regression with implicit and explicit attitudes predicting criterion bias revealed that implicit attitudes ($b = 0.08$, $t(201) = 2.03$, $p = .044$) but not explicit attitudes ($b = 0.03$, $t(201) = 0.91$, $p = .365$) reliably predicted differences in response criterion. A simultaneous linear regression predicting criterion bias from explicit attitudes, implicit attitudes, and perceived and desired performance suggested that implicit attitudes ($b = 0.17$, $t(199) = 2.16$, $p = .032$), perceived performance ($b = 0.12$, $t(199) = 2.57$, $p = .011$), and desired performance ($b = 0.17$, $t(199) = 2.58$, $p = .011$) contributed uniquely. Explicit attitudes ($b = -0.01$, $t(199) = -0.34$, $p = .735$) were not a reliable unique predictor of criterion bias. These variables accounted for 13.7% of the variance in criterion bias.

[4] Criterion and sensitivity were calculated in the same manner as Correll et al. (2007). Criterion = $-0.5 * (zFA + zH)$; Sensitivity = $zH - zFA$. FA is the percentage of false alarms and H the proportion of hits. The z operator represents standardized scores. To find computable z scores, FA and H were given a minimum of $1 / (2n)$ and maximum of $1 - (1 / 2n)$, where n = number of trials for each social group.

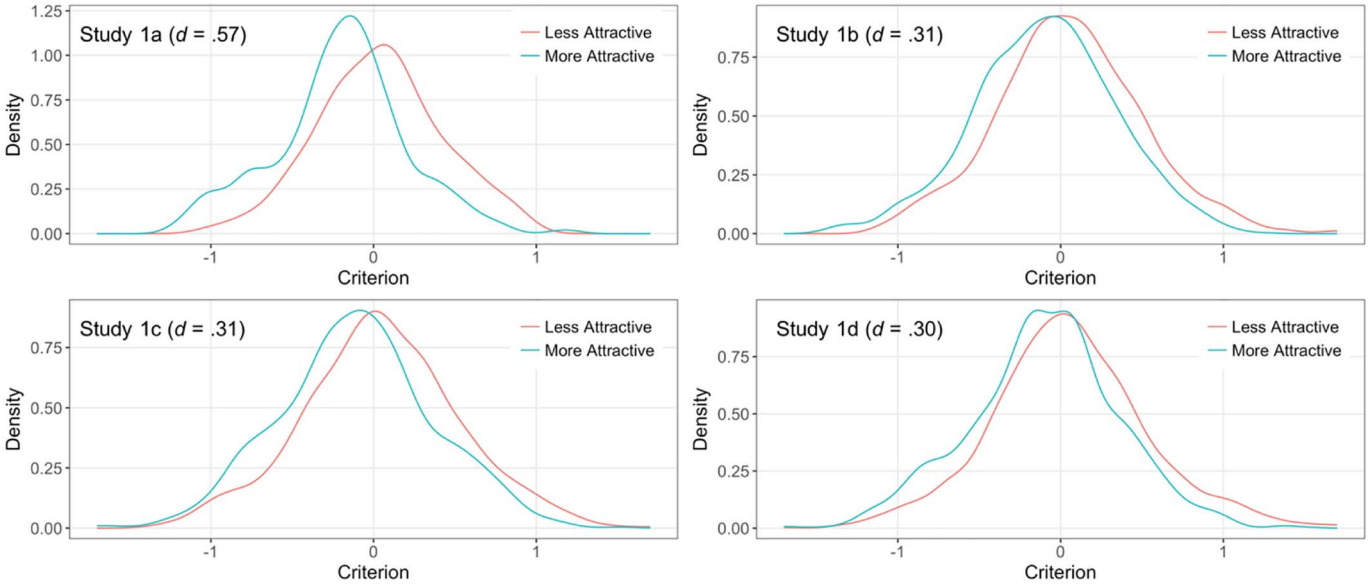

Fig. 1. Density plots of criterion towards more and less physically attractive profiles in Studies 1a–1d. The Cohen's *d* effect size among all eligible participants comparing the two criterion values is also reported.

2.3. Discussion

Participants had a lower criterion for more than less attractive applicants. Perceived performance, desired performance, and implicit but not explicit attitudes were reliably related to the criterion bias.

We replicated these effects in three large online samples, two from Project Implicit (Studies 1b and 1c), and one from an online sampling firm (Study 1d).[5] See Table 1 for sample sizes, descriptive and test statistics for criterion and sensitivity and Table 2 for correlations of criterion bias with perceived performance, desired performance, explicit attitudes and implicit attitudes. The online supplement includes study pre-registrations and full methods and results sections. Participants again had a lower criterion for more versus less physically attractive applicants (average $d = 0.31$) but no reliable differences in sensitivity (average $d = 0.01$).

While behavior in Studies 1a-1d was related to explicit attitudes as well as desired and perceived task performance, we investigated whether the attractiveness bias in criterion existed even among participants who reported either having no explicit preference, not wanting to show bias, or having shown no bias, pre-registering these analyses for Studies 1c-1d. Participants who stated that they treated more attractive and less attractive applicants equally (Study 1c: 75%, Study 1d: 82%) had lower criterion for more versus less attractive applicants (Study 1c: $t(622) = 3.95$, $p < .001$, $d = 0.16$, Study 1d: $t(1118) = 8.01$, $p < .001$, $d = 0.24$). Participants who stated that they wanted to treat more attractive and less attractive applicants equally (Study 1c: 90%, Study 1d: 86%) also had lower criterion for more versus less attractive applicants (Study 1c: $t(735) = 7.12$, $p < .001$, $d = 0.26$, Study 1d: $t(1170) = 9.30$, $p < .001$, $d = 0.27$). Finally, participants who stated that they had no explicit preference for more versus less attractive people (Study 1c: 39%, Study 1d: 56%) had lower criterion for more versus less attractive applicants (Study 1c: $t(316) = 3.53$, $p < .001$, $d = 0.20$, Study 1d: $t(765) = 6.64$, $p < .001$, $d = 0.24$).

Participants who reported wanting to show no bias on the task, having shown no bias on the task, and holding no preferences between more attractive and less attractive people all had lower criterion for more attractive relative to less attractive applicants. However, explicit preferences, perceived performance, and desired performance were reliable predictors of criterion bias. A lack of favoritism, a desire to show no favoritism, and a perception of having shown no favoritism are related to reduced criterion bias, but such preferences, desires, and perceptions were not sufficient to eradicate the judgment bias.

In Studies 2a & 2b, we extended the JBT to another well-known bias: ingroup favoritism. Favoritism towards one's ingroup has been a focus in psychological research since Sumner (1906) designated the term ethnocentrism to describe positively evaluating one's ingroup. In work on social identity theory, Tajfel and Turner (1979) found that membership to even arbitrarily defined social groups greatly determined self-categorization and influenced decisions in favor of one's ingroup with one meta-analysis finding an average ingroup bias effect of Cohen's $d = 0.36$ (Mullen, Brown, & Smith, 1992). We applied the JBT to investigating ingroup bias by now including information about whether the applicant came from the undergraduate participant's university or an academically similar university.

3. Study 2a

3.1. Methods

3.1.1. Participants

We sought to collect 160 participants. We arrived at this number by estimating that the same percentage of respondents in Study 2a would report showing no bias as Study 1a (56%; 89 out of 160). 89 participants would provide > 95% power at detecting an effect of differences in criterion equal to the size of that among participants in Study 1a who reported showing no bias on the JBT ($t(114) = 4.18$, $p < .001$, $d = 0.39$). Due to overscheduling, our sample was slightly larger: 169 University of Virginia undergraduates ($M_{age} = 18.6$, $SD = 0.89$; 62.7% women; 64.5% White) completed the study for partial course credit.

3.1.2. Design

The study used was a within-subjects design consisting of two levels of applicant school: UVA or University of North Carolina (UNC).

3.1.3. Procedure

Participants were run in groups of one to four with participants completing the study on computers in individual carrels without interaction among participants. Participants completed measures in the

[5] Participants in Studies 1b-1d completed a four-block, good-focal Brief Implicit Association Test (Sriram & Greenwald, 2009) measuring evaluations of more vs. less attractive people. See online supplement for procedure and scoring details.

Table 1
Descriptive and test statistics for Studies 1a–1d.

Study	*N*	Accept rate	Accuracy	More Attr. *c*	Less Attr. *c*	Comparison *d*	More Attr. d′	Less Attr. d′	Comparison *d*
1a	204	52.7%	70.4%	−0.20 (0.41)	0.04 (0.38)	0.57	1.18 (0.63)	1.25 (0.63)	−0.09
1b	1670	50.8%	66.6%	−0.09 (0.44)	0.06 (0.45)	0.31	0.98 (0.63)	0.98 (0.63)	0.001
1c	959	51.1%	66.0%	−0.11 (0.47)	0.05 (0.48)	0.31	0.96 (0.66)	0.95 (0.70)	0.01
1d	1542	50.9%	64.2%	−0.10 (0.46)	0.05 (0.48)	0.30	0.84 (0.66)	0.83 (0.67)	0.02

Note: More Attr = more attractive applicants. Less Attr = less attractive applicants. *c* = criterion. d′ = sensitivity *d* = Cohen's *d* effect size.

Table 2
Correlations between criterion bias with performance and attitude measures.

	Perc. Performance	Des. Performance	Exp. Attitudes	Imp. Attitudes
Study 1a	0.30 [0.17, 0.42]	0.29 [0.16, 0.41]	0.09 [−0.05, 0.22]	0.15 [0.02, 0.28]
Study 1b	0.28 [0.23, 0.33]	0.06 [0.003, 0.10]	0.12 [0.07, 0.17]	0.10 [0.05, 0.16]
Study 1c	0.18 [0.11, 0.24]	0.13 [0.07, 0.20]	0.12 [0.06, 0.19]	0.13 [0.06, 0.20]
Study 1d	0.15 [0.10, 0.20]	0.13 [0.08, 0.18]	0.13 [0.08, 0.18]	0.04 [−0.05, 0.16]

Note: Values are Pearson correlation coefficients and 95% confidence intervals. Perc. Performance = perceived performance. Des. Performance = desired performance. Exp. Attitudes = Explicit attitudes. Imp. Attitudes = BIAT D scores. Correlations with implicit attitudes exclude participants with > 10% of BIAT responses faster than 300 ms (Nosek, Bar-Anan, Sriram, Axt, & Greenwald, 2014).

following order: academic decision making task, explicit and implicit attitudes measures in a randomized order, and demographics. See https://osf.io/vuek8/ for the study's pre-registration.

3.1.3.1. Academic decision-making task. Participants completed the same task as in Study 1a except instead of having each application randomly paired with a photograph, each application was randomly paired with a UVA logo or a UNC logo. Logos associated with each application were randomly paired such that profiles from each level of qualification were matched with 16 logos from each school. We also changed the study instructions, noting that applicants would be coming from both UVA and UNC. We told participants that given the schools' similarity, they should consider both schools to be "equally rigorous."

3.1.3.2. Perceptions of performance. Participants answered two items about perceived and desired performance. These items were the same as Studies 1a–1d, now using the terms "UVA" and "UNC" instead of "more physically attractive" and "less physically attractive".

3.1.3.3. Explicit preferences. Participants reported their preference for UNC and UVA students using the same item from Studies 1a–1d, now using the terms "UVA students" and "UNC students" instead of "more physically attractive people" and "less physically attractive people".

3.1.3.4. Demographics. Participants completed the same seven-item demographics questionnaire as Study 1a. We only analyzed the gender, age and race items.

3.1.3.5. Implicit preferences. Participants completed an IAT measuring the strength of the association between the concepts "Good" and "Bad" and the categories "UVA" and "UNC". Images related to each school (logos, seals) were used as stimuli.

3.2. Results

In Study 2a, we first examined differences in criterion among all eligible participants, and then separately among those participants who reported a desire of being unbiased, a perception of having been unbiased, and no explicit preferences for UVA versus UNC students. We then analyzed how biases in criterion related to explicit attitudes, implicit attitudes, perceived performance and desired performance.

One participant was excluded from analyses for accepting < 20% or > 80% of the applicants, or for accepting or rejecting all applicants from either school. No participants were excluded for having > 10% of IAT response trials faster than 300 ms.

Accuracy on the task was 69.5% (SD = 7.7). The average acceptance rate was close to 50% (M = 52.2%, SD = 10.1). Participants required 5.31 min on average (SD = 1.47) to complete the task.

Of primary interest was the difference in criterion for UVA compared to UNC applicants. UVA applicants (M = −0.14, SD = 0.38) received a lower criterion than UNC applicants (M = 0.01, SD = 0.36), t(167) = 5.31, p < .001, d = 0.41, 95% C.I. [0.25, 0.57]. See Fig. 2 for density plots of criterion towards ingroup and outgroup members for Studies 2a–4.

Unlike earlier studies, internal reliability of criterion for UVA applicants (α = 0.59) was higher than reliability of criterion for UNC applicants (α = 0.49), and reliability of the criterion difference score was particularly low, α = 0.14. In an exploratory analysis, sensitivity was similar between UVA (M = 1.16, SD = 0.65) and UNC (M = 1.13, SD = 0.61) applicants, d = 0.05, 95% CI [−0.10, 0.20].

One hundred and eight participants (64.3%) stated that they treated UVA and UNC applicants equally. Among them, UVA applicants (M = −0.15, SD = 0.39) received a lower criterion than UNC applicants (M = −0.05, SD = 0.36), t(107) = 3.01, p = .003, d = 0.29, 95% C.I. [0.10, 0.48]. One hundred and thirty-eight participants (82.1%) stated they wanted to treat UVA and UNC applicants equally. Among them, UVA applicants (M = −0.12, SD = 0.38) received a lower criterion than UNC applicants (M = 0.004, SD = 0.36), t(137) = 3.93, p < .001, d = 0.33, 95% C.I. [0.16, 0.51]. Forty-three participants (25.6%) stated that they had no preference for UVA or UNC students. Among them, UVA applicants (M = −0.12, SD = 0.33) received a lower criterion than UNC applicants (M = −0.06, SD = 0.28), but this comparison was not statistically reliable, t(42) = 1.34, p = .188, d = 0.20, 95% C.I. [−0.10, 0.51].

3.2.1. Predicting criterion bias

IAT D scores (M = 0.43, SD = 0.35, d = 1.23) and the explicit preference (M = 1.18, SD = 0.92, d = 1.28) item indicated pro-UVA attitudes.

We computed the same criterion difference score as Studies 1a–1d, such that higher values meant lower criterion for UVA relative to UNC applicants. This difference score was not reliably correlated with IAT D scores (r(168) = −0.02, p = .801, 95% C.I. [−0.17, 0.13]), but was positively and reliably correlated with perceptions of performance (r(168) = 0.30, p < .001, 95% C.I. [0.16, 0.43]), desired performance (r(168) = 0.26, p = .001, 95% C.I. [0.11, 0.39]), and explicit preference (r(168) = 0.23, p = .002, 95% C.I. [0.08, 0.37].

A simultaneous linear regression predicting criterion bias from implicit and explicit attitudes revealed that that explicit (b = 0.09, t(165) = 3.09, p = .002) but not implicit attitudes (b = −0.04, t(165) = −0.45, p = .652) were reliable predictors of criterion bias. Finally, a simultaneous linear regression predicting criterion bias from implicit attitudes, explicit attitudes, perceived performance and desired

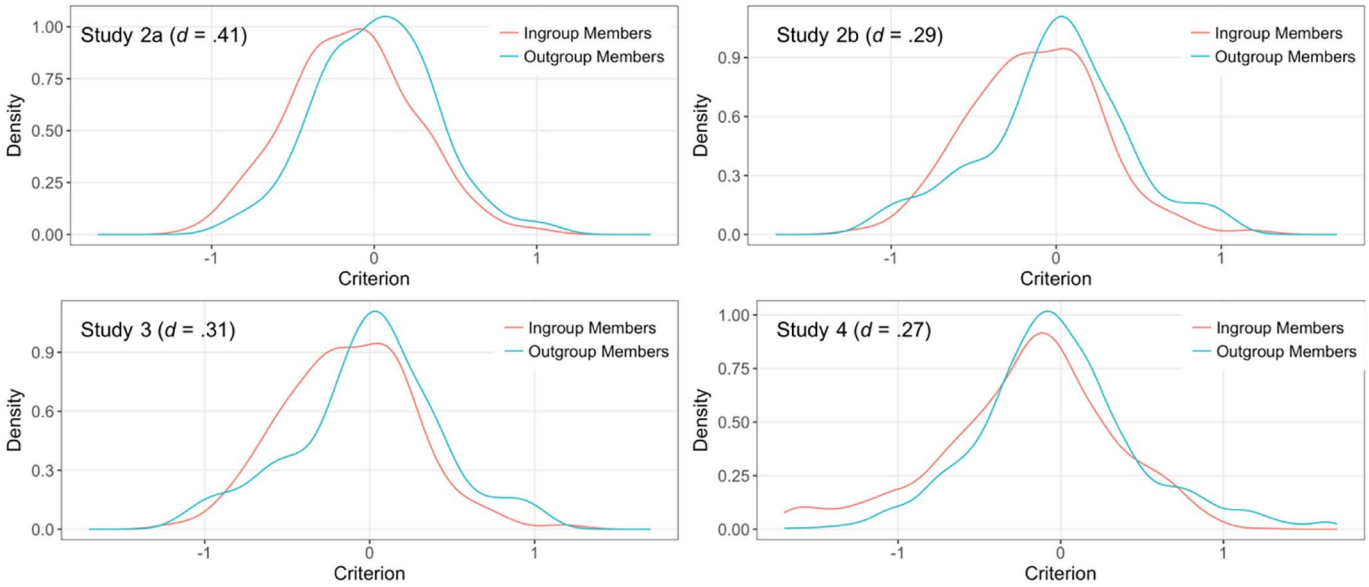

Fig. 2. Density plots of criterion towards ingroup and outgroup profiles in Studies 2a–4. In Study 2a, ingroup members are from UVA and outgroup members are from UNC, with the reverse in Study 2b. In Study 3, ingroup members are from one's own political party and outgroup members are from the other political party (collapsing across self-reported Democrats and Republicans). In Study 4, ingroup members are White profiles and outgroup members are non-White (Black and Hispanic) profiles. The Cohen's *d* effect size among all eligible participants comparing the two criterion values is also reported.

performance revealed that explicit attitudes ($b = 0.07$, $t(163) = 2.30$, $p = .023$), and perceived performance ($b = 0.11$, $t(163) = 2.32$, $p = .021$) contributed uniquely. Implicit attitudes ($b = -0.04$, $t(163) = -0.55$, $p = .584$) and desired performance ($b = 0.10$, $t(163) = 1.80$, $p = .074$) were not unique predictors. These variables accounted for 13.7% of the variance in criterion bias.

3.3. Discussion

As in Studies 1a–1d, participants displayed a criterion bias, with a lower criterion for applicants from one's own university versus another university. Again, perceived and desired task performance were related to levels of criterion bias, but bias was present even among those who reported showing no bias and who reported wanting to show no bias, again suggesting that perceived and desired performance are related to judgment biases but are not sufficient to account for such biases.

In Study 2b, we sought to replicate the effect of ingroup bias in criterion among UNC students. 151 UNC undergraduates completed the same measures as the UVA undergraduates in Study 2a. We sought to recruit at least 150 participants, estimating that the same percentage of participants would report not showing bias on the task as in Study 2a (64.3%; 96 out of 150). These 96 participants would provide 80% for detecting the size of the criterion bias displayed by participants in Study 2a who reported showing no bias on the task ($t(107) = 3.01$, $p = .003$, $d = 0.29$). See https://osf.io/2wvdm/ for the study's pre-registration and the online supplement for full methods and results. We replicated ingroup favoritism in criterion. UNC applicants received a lower criterion than UVA applicants ($t(142) = 3.46$, $p = .001$, $d = 0.29$) and this persisted among those stating they wanted to treat UNC and UVA applicants equally (81.7% of sample, $t(115) = 2.14$, $p = .035$, $d = 0.20$), but not among participants stating they treated UNC and UVA applicants equally (69.9%, $t(99) = 0.87$, $p = .385$, $d = 0.09$) or among those reporting no preference for UVA or UNC people (38.7%, $t(54) = -0.78$, $p = .440$, $d = -0.10$). The internal reliability of the criterion for UVA applicants ($\alpha = 0.72$) was higher than the reliability of criterion for UNC applicants ($\alpha = 0.53$). The criterion difference score reliability was $\alpha = 0.36$.

In Study 3, we tested whether ingroup biases in criterion would also be present for another social category: political orientation.

4. Study 3

4.1. Methods

4.1.1. Participants

We sought to collect at least 300 participants self-identified Democrats, Republicans, and Independents from the Project Implicit research pool. The 300 participants from each group provided > 95% power for detecting an effect the same size of the criterion bias found in Study 1b, which used participants from the same source ($t(1535) = 12.16$, $p < .001$, $d = 0.31$).

As studies on Project Implicit at the time were taken down on fixed days and conservatives are less represented in the pool than liberals, the final sample was larger: 1621 participants (Democrats $n = 688$, Republicans $n = 368$, Independents $n = 565$) volunteered, consented, and provided data.

We limited data collection to American citizens and residents over the age of 18. Participants provided this demographic information when first registering for the research pool. Among those who provided data, 64.4% were female, 78.3% were White, and the mean age was 39.5 ($SD = 14.6$). Sample sizes vary among tests due to missing data.

4.1.2. Procedure

The study session consisted of three components completed in the following order: the academic decision-making task, a survey about task performance and explicit attitudes, a survey about political orientation, and a measure of implicit identification with Democrats and Republicans. See https://osf.io/h7kqp/ for the study's pre-registration.

4.1.2.1. Academic decision-making task. Participants completed the same academic decision-making task as in previous studies with the following changes. First, participants only saw 16 applicants during the preview phase, four applicants for each qualification level and political orientation combination. Second, applicants were presented with the participant's political orientation (Democratic or Republican) and another irrelevant piece of information, number of siblings (1–3). We added the sibling information to make it less obvious that political orientation was the variable of interest. Participants were randomly assigned to one out of 18 orders. Across orders, each application was

equally likely to be described as a Democratic or Republican. Within each order, the 16 applicants belonging to each level of qualification level and political orientation had the same number of applicants with 1, 2, or 3 siblings.

4.1.2.2. Perception of performance and explicit preferences. Participants completed the same three items about perceived performance, desired performance and explicit preferences used in previous studies, updated to assess perceived and desired favoritism towards Democratic or Republican applicants and explicit preferences for Democrats relative to Republicans.

4.1.2.3. Political attitudes and identification. Participants completed a five-item survey about political attitudes (Hawkins & Nosek, 2012). First, participants responded to the question, "In general, how liberal or conservative are you on social issues (e.g., abortion, gay marriage, gun control)?". Next, participants responded to the question, "In general, how liberal or conservative are you on economic issues (e.g., free market policies, taxation)?" These questions had a seven-point response scale ranging from "Strongly liberal" to "Strongly conservative".

Next, participants reported their political identification, selecting from the following options: Democrat; Republican; Independent- I do not identify with any party; Libertarian; Green; Other; Don't Know. We used responses to this question to classify participants as Democrat, Republican, or Independent. If participants selected either "Democrat" or "Republican", they answered a follow-up question asking how strongly they identify with their selected party (slightly, moderately, or strongly). If participants selected "Independent", they answered a follow-up question of, "If you had to choose, between Democrats and Republicans, how would you identify your political affiliation?", using a seven-point response scale ranging from Strongly Republican to Strongly Democrat, with a neutral midpoint of "Independent."

4.1.2.4. Implicit identification. Implicit identification was measured using a four-block, self-focal BIAT. The targets were "Democrats" and "Republicans", with stimuli consisting of "Democrat words" (*Democrat*, *Barack Obama*, *Left Wing*, *Liberal*) and "Republican words" (*Conservative*, *Right Wing*, *George Bush*, *Republican*). The categories were "Self words" (*Mine*, *Myself*, *Self*, *I*, *My*) and "Other words" (*They*, *Them*, *Their*, *Theirs*, *Other*). Participants were randomly assigned to complete one of the two possible orders. BIAT responses were scored such that more positive scores reflected stronger associations between the self and Democrat. Procedural details are available in supplementary information.

4.2. Results

In Study 3, we first examined differences in criterion among Republican and Democrat participants separately. We then compared whether the size of the ingroup bias in criterion differed between Democrats and Republicans. Next, we divided self-identified Independents into implicit-Democrats and implicit-Republicans based on their BIAT results, and analyzed whether biases in criterion on the JBT emerged within each group. We then examined biases in criterion among Democrats, Republicans and Independents who reported a desire to be unbiased, a perception of having been unbiased, and no explicit preferences between Democrats and Republicans. Finally, we analyzed how biases in criterion related to implicit attitudes, explicit attitudes, perceived performance and desired performance.

Participants were excluded from analysis for accepting < 20% or > 80% of the applicants, or for accepting or rejecting every Democratic or Republican applicant. 107 participants (7.2%) were excluded based on these criteria. 34 additional participants (2.6% of those completing the BIAT) were excluded from analyses involving the BIAT for having > 10% of responses faster than 300 ms.

Accuracy on the task was 68.0% (SD = 8.0). The average acceptance rate was close to 50% (M = 52.1%, SD = 12.2). Participants required 5.52 min on average (SD = 3.09) to complete the task.

4.2.1. Criterion bias in decision-making

We analyzed criterion biases separately for self-identified Democrats and Republicans. Among Democratic participants, Democratic applicants (M = −0.20, SD = 0.49) received a lower criterion than Republican applicants (M = 0.004, SD = 0.50), t(640) = 8.61, p < .001, d = 0.34, 95% C.I. [0.26, 0.42]. Conversely, among Republican participants, Republican applicants (M = −0.09, SD = 0.47) received a lower criterion than Democratic applicants (M = 0.03, SD = 0.45), t(337) = 4.65, p < .001, d = 0.25, 95% C.I. [0.14, 0.36]. Across all participants, internal reliability of the criterion for Democratic (α = 0.74) and Republican (α = 0.74) applicants were comparable. The internal reliability of the criterion difference score was α = 0.61.

Exploratory analyses showed similar sensitivity for Democratic (Democratic participants: M = 1.10, SD = 0.60; Republican participants: M = 1.08, SD = 0.62) versus Republican (Democratic participants: M = 1.05, SD = 0.64; Republican participants: M = 1.01, SD = 0.67) applicants (Democratic participants: d = 0.07, 95% CI [−0.01, 0.15]; Republican participants: d = 0.09, 95% CI [−0.01, 0.20]).

For Democratic and Republican participants, we computed an ingroup bias criterion score (Other party criterion – own party criterion). Democratic participants showed a slightly larger ingroup criterion bias (M = 0.21, SD = 0.61) than Republican participants (M = 0.12, SD = 0.47), t(977) = 2.29, p = .022, d = 0.16, 95% C.I. [0.02, 0.29].

Next, we divided Independent participants into "implicitly identified Democrats" (n = 284) and "implicitly identified Republicans" (n = 181) based on their BIAT D scores (positive D scores categorized as implicit Democrats, negative D scores categorized as implicit Republicans). Implicitly identified Democrats had a lower criterion for Democratic (M = −0.11, SD = 0.45) than Republican applicants (M = 0.01, SD = 0.43), t(283) = 4.92, p < .001, d = 0.29, 95% C.I. [0.17, 0.41]. However, implicitly identified Republicans showed no reliable difference in criterion for Democratic (M = −0.001, SD = 0.46) versus Republican applicants (M = 0.02, SD = 0.48), t(280) = 0.73, p = .468, d = 0.06, 95% C.I. [−0.09, 0.20].

In exploratory analyses, sensitivity was similar among Independents who implicitly identified as Democrats for Democratic (M = 1.13, SD = 0.62) versus Republican applicants (M = 1.13, SD = 0.60), d = 0.01, 95% CI [−0.11, 0.12]. However, among Independents who implicitly identified with Republicans, sensitivity was higher for Republican (M = 1.20, SD = 0.64) than Democratic (M = 1.02, SD = 0.63) applicants, d = 0.26, 95% CI [0.11, 0.40].

4.2.2. Criterion bias and explicit attitudes, perceptions of performance, and desired performance

419 Democrats (66.3%) stated that they treated Democratic and Republican applicants equally. Among them, Democratic applicants (M = −0.12, SD = 0.43) received a lower criterion than Republican applicants (M = −0.07, SD = 0.45), t(418) = 2.91, p = .004, d = 0.14, 95% C.I. [0.05, 0.24]. 266 Republicans (79.4%) stated that they treated Democratic and Republican applicants equally. Among them, Republican applicants (M = −0.05, SD = 0.44) received a slightly but non-significantly lower criterion than Democratic applicants (M = −0.01, SD = 0.42), t(265) = 1.82, p = .070, d = 0.11, 95% C.I. [−0.01, 0.23]. 410 Independents (79.5%) stated that they treated Democratic and Republican applicants equally. Among them, Democratic applicants (M = −0.06, SD = 0.45) received a lower criterion than Republican applicants (M = −0.01, SD = 0.45), t(409) = 2.73, p = .007, d = 0.13, 95% C.I. [0.04, 0.23].

516 Democrats (81.3%) stated that they wanted to treat Democratic and Republican applicants equally. Among them, Democratic applicants (M = −0.16, SD = 0.44) received a lower criterion than

Republican applicants ($M = -0.03$, $SD = 0.46$), $t(515) = 6.47$, $p < .001$, $d = 0.28$, 95% C.I. [0.20, 0.37]. 284 Republicans (84.3%) stated that they wanted to treat Democratic and Republican applicants equally. Among them, Republican applicants ($M = -0.07$, $SD = 0.44$) received a lower criterion than Democratic applicants ($M = 0.001$, $SD = 0.43$), $t(283) = 3.33$, $p = .001$, $d = 0.20$, 95% C.I. [0.08, 0.31]. 456 Independents (88.5%) stated that they wanted to treat Democratic and Republican applicants equally. Among them, Democratic applicants ($M = -0.07$, $SD = 0.45$) received a lower criterion than Republican applicants ($M = 0.01$, $SD = 0.45$), $t(455) = 4.08$, $p < .001$, $d = 0.19$, 95% C.I. [0.10, 0.28].

81 Democrats (12.8%) stated that they had no explicit preference for Democrats vs. Republicans. Among them, there was no reliable difference in criterion for Democratic applicants ($M = -0.21$, $SD = 0.35$) versus Republican applicants ($M = -0.17$, $SD = 0.41$), $t(80) = 1.10$, $p = .276$, $d = 0.12$, 95% C.I. [−0.10, 0.34]. 103 Republicans (30.5%) stated that they had no preference for Democrats vs. Republicans. Among them, Republican applicants ($M = -0.09$, $SD = 0.47$) received a non-significantly lower criterion than Democratic applicants ($M = -0.03$, $SD = 0.46$), $t(102) = 1.94$, $p = .055$, $d = 0.19$, 95% C.I. [−0.004, 0.39]. 197 Independents (38.1%) stated that they had no preference for Democrats vs. Republicans. Among them, there was no reliable difference in criterion for Democratic applicants ($M = -0.04$, $SD = 0.46$) versus Republican applicants ($M = -0.04$, $SD = 0.48$), $t(196) = 0.07$, $p = .946$, $d = 0.01$, 95% C.I. [−0.13, 0.15].

4.2.3. Predicting criterion bias

Among Democrats, BIAT D scores ($M = 0.57$, $SD = 0.47$, $d = 1.21$) and the explicit preference item ($M = 1.96$, $SD = 1.08$, $d = 1.81$) indicated pro-Democrat attitudes. Among Republicans, implicit ($M = -0.55$, $SD = 0.45$, $d = -1.22$) and explicit ($M = -1.13$, $SD = 1.08$, $d = -1.04$) attitudes favored Republicans. Among Independents, implicit ($M = 0.15$, $SD = 0.54$, $d = 0.28$) and explicit ($M = 0.67$, $SD = 1.46$, $d = 0.46$) attitudes favored Democrats.

We computed another criterion difference score, such that higher values meant lower criterion for Democratic relative to Republican applicants. The difference score was positively and reliably correlated with BIAT D scores ($r(1294) = 0.22$, $p < .001$, 95% C.I. [0.17, 0.27]), explicit preferences for Democrats vs. Republicans ($r(1489) = 0.31$, $p < .001$, 95% C.I. [0.26, 0.36]), perceptions of performance ($r(1483) = 0.47$, $p < .001$, 95% C.I. [0.43, 0.51]), and desired performance ($r(1487) = 0.40$, $p < .001$, 95% C.I. [0.35, 0.44]).

A simultaneous linear regression with implicit and explicit attitudes predicting criterion bias revealed that explicit ($b = 0.09$, $t(1281) = 8.35$, $p < .001$) but not implicit attitudes ($b = 0.02$, $t(1281) = 0.77$, $p = .442$) reliably predicted differences in response criterion. Another simultaneous linear regression including explicit attitudes, implicit attitudes, and perceived and desired performance revealed that explicit attitudes ($b = 0.03$, $t(1269) = 2.60$, $p = .010$) perceived performance ($b = 0.22$, $t(1269) = 11.73$, $p < .001$) and desired performance ($b = 0.15$, $t(1269) = 6.44$, $p < .001$) contributed uniquely. Implicit attitudes ($b = 0.02$, $t(1269) = 0.57$, $p = .566$) were not a reliable, unique predictors of criterion bias. These variables accounted for 24.4% of the variance in criterion bias.

4.3. Discussion

Republican and Democratic participants had a lower criterion for members of their own relative to the rival political party. Independents who implicitly identified as Democrats had lower criterion for Democratic than Republican applicants, though there were no reliable differences in criterion among Independents implicitly identifying as Republicans. Criterion biases largely persisted among Independent, Democratic, and Republican participants who indicated not wanting to show favoritism on the task and not showing favoritism on the task. However, perhaps understandably, few participants reported no explicit preferences between Democrats and Republicans (13% of Democrats, 31% of Republicans and 38% of Independents), and such participants did not show reliable evidence of criterion bias.

Studies 1–3 focused on using the JBT within an academic context. In Study 4, we highlight the flexible nature of the JBT by investigating bias in another social category (race), context (dating), stimulus design (six qualifications instead of four) and using three target social groups. In Study 4, White participants evaluated White, Hispanic, and Black profiles for a hypothetical dating website.

5. Study 4

5.1. Participants

Since this study was investigating criterion biases in a new domain, we did not rely on past studies to calculate our sample size. We sought to collect 800 participants from the Project Implicit research pool, which provided > 95% power for detecting a small within-subjects effect of Cohen's $d = 0.20$. As studies on Project Implicit at the time were taken down on fixed days, the final sample was larger: 1100 participants volunteered, consented, and provided data. We limited data collection to White participants < 30 years old because Whites were the most plentiful available sample, and younger participants maximized the relevance of the dating context. Among those who provided data, 64.4% were female and the mean age was 21.5 ($SD = 3.3$).

5.2. Procedure

The study session consisted of three components completed in the following order: the interracial dating decision making task, a survey about task performance, racial attitudes and dating preferences, and a measure of implicit attitudes towards White, Black, and Hispanic people. See https://osf.io/asgfq/ for the study's pre-registration.

5.2.1. Interracial rating decision-making task

Participants completed an interracial JBT similar to the academic JBT. Participants were first instructed that they would accept and reject profiles for an online dating site, and it was their task to accept applicants they would consider dating and reject those they would not consider dating. Next, in the viewing phase, participants passively viewed the 60 profiles for 1 s each. In the selection phase, participants made accept or reject decisions on each profile.

Each profile came with six pieces of information: attitude similarity (Scale of 1–10), social similarity (1 − 10), intelligence (1–4), openness (1–4), dependability (poor, fair, good, excellent), and sense of humor (poor, fair, good, excellent). Characteristics were represented with three different scales to reduce the likelihood that participants would use a simple decision rule (e.g., adding up the numbers). Participants were told to weigh each piece of information equally when evaluating profiles.

We varied the qualifications to create 60 unique profiles, 30 *more qualified* and 30 *less qualified*. To determine qualification, the six pieces of applicant information were converted to a scale with a maximum score of 4.[6] Intelligence and openness were already in this form, and we converted dependability and sense of humor (poor = 1, fair = 2, good = 3, excellent = 4) and attitude and social similarity (dividing by 2.5). *Less qualified* profiles summed to 19.5 and *more qualified* profiles

[6] Since dependability and sense of humor had whole-number values, the qualification scores could not have the same standard deviation across profiles while also producing 60 unique combinations. Profiles were made to have similar standard deviations between attitude similarity ($SD = 0.33$), social similarity ($SD = 0.36$), intelligence ($SD = 0.35$), and openness ($SD = 0.35$), as well as between dependability ($SD = 0.50$) and sense of humor ($SD = 0.50$).

summed to 21.

Profiles were also presented with the demographic information of race (White, Black, or Hispanic) and number of siblings (1, 2 or 3). Within each qualification level, 10 profiles were White, 10 were Black, and 10 were Hispanic. Within each combination of race and qualification level, four profiles had one sibling, four profiles had two siblings, and two profiles had three siblings. Participants were randomly assigned to one out of 18 orders. Across the 18 orders, each profile was equally likely to be described as White, Black, or Hispanic.

5.2.2. Perception of performance and racial and dating preferences

Participants completed 11 items about their racial attitudes, perceptions of performance and desired performance on the task, and dating preferences.

Participants completed three items related to perceived performance, one for performance towards White vs. Hispanics, one for Blacks vs. Whites, one for Hispanics vs. Blacks. Perceived performance was measured by the item, "Which statement best describes your performance on the task towards X and Y people?" (−3 = I was much more likely to accept profiles of X people than profiles of Y people, +3 = I was much more likely to accept profiles of Y people than profiles of X people). Participants completed the same three items for desired performance, measured by the item, "Which statement best describes how you wanted to perform on the task towards X and Y people?" (1 = I wanted to be much more likely to accept profiles of X people than profiles of Y people, 7 = I wanted to be much more likely to accept profiles of Y people than profiles of X people).

Participants then completed three items assessing preferences for White vs. Hispanic, White vs. Black and Hispanic vs. Black people using the same wording as in previous studies.

Next, participants responded to the item, "To what extent do you prefer to date people of your own race compared to people of other races?" (−3 = I strongly prefer dating people of other races compared to people of my own race, +3 = I strongly prefer dating people of my own race compared to people of other races). Finally, participants reported their current relationship status (Single, dating, or married).

5.2.3. Implicit attitudes

Participants completed a seven-block, good-focal Multi-Category Implicit Association Test (MC-IAT; Axt, Ebersole, & Nosek, 2014) measuring evaluations of White, Black and Hispanic people. Participants were randomly assigned to one of 12 MC-IAT orders. MC-IAT responses were scored by the *D* algorithm (Nosek et al., 2014). Procedural details are available in the online supplement.

5.3. Results

In Study 4, we first examined racial biases in criterion among all eligible participants, and then separately among participants who reported a desire of being unbiased, a perception of having been unbiased, or no preferences in racial attitudes. We then analyzed how biases in criterion for White vs. non-White profiles related to explicit attitudes, implicit attitudes, perceived performance and desired performance.

Ninety-nine participants were excluded from analysis for accepting < 20% or > 80% of the applicants on the decision-making task (9.0%).[7] Twenty-five additional participants were excluded from analyses involving the MC-IAT for having > 10% of MC-IAT trial responses < 300 ms (Nosek et al., 2014).

Accuracy on the task was 66.3% (SD = 9.9). The average acceptance rate was close to 50% (M = 52.2%, SD = 12.0). Participants required 5.80 min on average (SD = 2.86) to complete the task.

5.3.1. Bias in response criterion

Of primary interest was the difference in criterion for own-race compared to other-race profiles. We analyzed the data comparing both White vs. Non-White profiles and comparing each race specifically. White profiles (M = −0.21, SD = 0.54) received a lower criterion than non-White profiles (M = 0.004, SD = 0.54), $t(1000) = 8.65$, $p < .001$, $d = 0.27$, 95% C.I. [0.21, 0.34]. White profiles received a lower criterion than Black profiles (M = 0.003, SD = 0.64), $t(1000) = 7.59$, $p < .001$, $d = 0.24$, 95% C.I. [0.18, 0.30], and a lower criterion than Hispanic profiles (M = −0.003, SD = 0.57), $t(1000) = 8.60$, $p < .001$, $d = 0.27$, 95% C.I. [0.21, 0.33]. There was no reliable difference in the decision criterion for Black and Hispanic profiles, $t(1000) = 0.29$, $p = .773$, $d = 0.01$, 95% CI [−0.05, 0.07]. The internal reliability of the criterion measure was comparable for Black ($\alpha = 0.79$), Hispanic ($\alpha = 0.72$) and White ($\alpha = 0.68$) profiles. The reliability of the criterion difference score for White vs. Non-White profiles was $\alpha = 0.78$.

In exploratory analyses, White profiles (M = 0.94, SD = 0.78) had slightly higher sensitivity than Non-White profiles (M = 0.89, SD = 0.67), $d = 0.07$, 95% CI [0.01, 0.13]. This pattern was smaller when looking specifically at the contrast between sensitivity for White and Black profiles, $d = 0.06$, 95% CI [−0.003, 0.12], and between White and Hispanic profiles, $t(1000) = 1.56$, $d = 0.05$. Sensitivity was similar between Black (M = 0.83, SD = 0.79) and Hispanic (M = 0.83, SD = 0.79) profiles, $d = 0.01$, 95% CI [−0.05, 0.07].

398 participants (39.8%) stated that they treated profiles of each race equally. However, these participants actually displayed a pro-Black criterion bias; White profiles (M = −0.08, SD = 0.44) received a higher criterion than non-White profiles (M = −0.14, SD = 0.43), $t(397) = 2.74$, $p = .006$, $d = 0.14$, 95% C.I. [0.04, 0.24]. White profiles also received a higher criterion than Black profiles (M = −0.18, SD = 0.48), $t(397) = 3.88$, $p < .001$, $d = 0.19$, 95% C.I. [0.09, 0.29]. The difference in criterion between White and Hispanic profiles was not reliable (M = −0.11, SD = 0.47), $t(397) = 1.22$, $p = .225$, $d = 0.06$, 95% CI [−0.04, 0.16], and Black profiles also received a lower criterion than Hispanic profiles, $t(397) = 3.22$, $p = .001$, $d = 0.16$, 95% C.I. [0.06, 0.26].

577 participants (57.7%) stated that they wanted to treat profiles of each race equally. These participants displayed a small anti-Hispanic bias in criterion. There was no reliable difference in criterion between White (M = −0.12, SD = 0.47) and non-White profiles (M = −0.09, SD = 0.43), $t(576) = 1.57$, $p = .118$, $d = 0.07$, 95% CI [−0.02, 0.15] and between White and Black profiles (M = −0.11, SD = 0.51), $t(576) = 0.59$, $p = .554$, $d = 0.02$, 95% CI [−0.06, 0.11]. However, White profiles received a lower criterion than Hispanic profiles (M = −0.06, SD = 0.48), $t(576) = 2.48$, $p = .013$, $d = 0.10$, 95% C.I. [0.02, 0.19], and Black profiles received a slightly lower criterion than Hispanic profiles, $t(576) = 2.05$, $p = .041$, $d = 0.09$, 95% C.I. [0.003, 0.17].

408 participants (40.8%) stated that they had no preferences between White, Hispanic, and Black people. These participants displayed little bias in criterion. There was no reliable difference in criterion between White profiles (M = −0.09, SD = 0.47) and non-White profiles (M = −0.09, SD = 0.48), $t(407) = 0.03$, $p = .978$, $d = 0.002$, 95% CI [−0.06, 0.06] and between White and Black profiles (M = −0.13, SD = 0.55), $t(407) = 1.17$, $p = .241$, $d = 0.06$, 95% CI [−0.04, 0.16] and between White and Hispanic profiles (M = −0.06, SD = 0.52), $t(407) = 1.13$, $p = .259$, $d = 0.06$, 95% CI [−0.04, 0.15]. However, Black profiles received a lower criterion than Hispanic profiles, $t(407) = 2.73$, $p = .007$, $d = 0.14$, 95% C.I. [0.04, 0.23].

We tested whether any of the above analyses were moderated by participant gender and relationship status. Only one of 16 gender moderation analyses was reliable at $p < .05$. In addition, only four of the 16 relationship status moderation analyses were reliable at the $p < .05$ level. All analyses are available in the online supplement.

[7] Unlike previous studies, participants who accepted or rejected all profiles from one race were not excluded from analyses, as we perceived it realistic that some participants would want to show racial preferences when evaluating potential romantic partners.

5.3.2. Predicting bias in criterion

One of the orders in the MC-IAT did not record one block of the White vs. Hispanic BIAT. This error was corrected during data collection, but participants assigned to that order have missing data for the White vs. Hispanic BIAT and the White and Hispanic aggregate MC-IAT scores.

The BIAT *D* within the MC-IAT scores indicated more positive associations for Whites vs. Hispanics ($M = 0.18$, $SD = 0.52$, $d = 0.35$) and Whites vs. Blacks ($M = 0.19$, $SD = 0.57$, $d = 0.33$), with neutral associations for Hispanics vs. Blacks ($M = 0.01$, $SD = 0.50$, $d = 0.02$). Descriptively, aggregate MC-IAT scores showed more positive associations for Whites ($M = 0.19$, $SD = 0.43$, $d = 0.44$) than Blacks ($M = -0.10$, $SD = 0.40$, $d = -0.25$) or Hispanics ($M = -0.08$, $SD = 0.37$, $d = -0.22$).

We computed a pro-White criterion difference score such that higher values meant lower criterion for White relative to non-White profiles. This difference score was positively correlated with aggregate MC-IAT evaluations of White people ($r(802) = 0.29$, $p < .001$, 95% C.I. [0.23, 0.35]), explicit preferences for White people ($r(951) = 0.41$, $p < .001$, 95% C.I. [0.36, 0.46]) perceptions of performance ($r(948) = 0.62$, $p < .001$, 95% C.I. [0.58, 0.65]), desired performance ($r(944) = 0.52$, $p < .001$, 95% C.I. [0.47, 0.56]), and attitudes towards interracial dating ($r(948) = 0.43$, $p < .001$, 95% C.I. [0.38, 0.48]). See Table 3 for a correlation matrix.

A simultaneous linear regression with implicit and explicit attitudes predicting the pro-White criterion bias revealed that implicit attitudes ($b = 0.33$, $t(795) = 5.56$, $p < .001$) and explicit attitudes ($b = 0.28$, $t(795) = 9.86$, $p < .001$) reliably predicted differences in response criterion. A simultaneous linear regression predicting criterion bias from explicit attitudes, implicit attitudes, perceived performance, desired performance and attitudes towards interracial dating revealed that explicit attitudes ($b = 0.01$, $t(782) = 0.28$, $p = .783$) were not reliable predictors of criterion bias, while implicit attitudes ($b = 0.19$, $t(782) = 3.81$, $p < .001$), perceived performance ($b = 0.32$, $t(782) = 11.40$, $p < .001$), desired performance ($b = 0.22$, $t(782) = 6.94$, $p < .001$) and attitudes towards interracial dating ($b = 0.07$, $t(782) = 3.64$, $p < .001$) contributed uniquely. These variables accounted for 45% of the pro-White criterion bias.

The online supplement contains correlation tables and analyses for the criterion specific to White vs. Black and White vs. Hispanic profiles, which were reliably correlated to measures of implicit attitudes, explicit attitudes, perceived performance, desired performance, and attitudes towards interracial dating (all r's > 0.18, all p's $< .001$).

5.4. Discussion

White participants on average showed a lower criterion for White than Black or Hispanic dating profiles, demonstrating a criterion bias in a new context (dating), towards new social categories (race), and with new stimuli (a six-factor profile assessing more abstract qualities like sense of humor). This criterion bias was related to implicit and explicit attitudes, as well as perceived performance, desired performance, and attitudes towards interracial dating.

However, unlike previous studies, ingroup favoritism in criterion did not emerge among participants who reported showing no bias on the task or wanting to show no bias on the task. Among participants who reported showing no bias on the task, there was actually a pro-Black criterion bias, a result that mirrors a similar pro-Black bias among White participants in an academic context (Axt, 2017; Axt, Ebersole, & Nosek, 2016). Among participants who reported wanting to show no bias on the task, Hispanic profiles received a higher criterion than Black or White profiles.

One potential reason for this subgroup's lack of ingroup bias was that much of the sample reported a preference for dating members of their own race. Nearly 61% of participants had at least a slight preference to date members of their own compared to another race, with 15% reporting an “extreme” preference to do so. This comfort with expressing racial preferences in dating partners may mean that reported perceptions or desires to behave fairly were better able to distinguish participants who genuinely thought they were or wanted to be fair from those who felt normative pressure to report a desire and perception of fairness. Indeed, the percentage of participants in Study 4 who reported a perception of being fair (41%) was considerably lower than Project Implicit samples completing the academic version of the JBT dealing with more attractive and less attractive applicants (73% in Study 1b; 75% in Study 1c) or Republicans and Democrats applicants (74% in Study 3). Likewise, only 57% of Study 4 participants reported a desire to be fair, compared to 88% in Study 1b, 90% in Study 1c, and 85% in Study 3. These results indicate that, if one intends to have a bias, that intention can be manifested with the JBT relatively straightforwardly, but evidence from the other studies indicate an asymmetry in controllability. If one intends to *not* have a bias, this intention may not be sufficient to avoid showing it on the JBT.

6. Study 5

Studies 1–4 show that the JBT effectively measures bias in social judgment, often among participants reporting a desire to show no bias on the task and a perception of having done so. In a final study, we investigated whether an intervention could reduce or eliminate such biases. A recent study suggests that evaluating applicants side-by-side reduces gender biases compared to evaluating one at a time (Bohnet, van Geen, & Bazerman, 2015). The ingenious concept is that joint evaluations make it easier to focus reasoning on comparing relevant criteria and reduces the unintended influence of irrelevant criteria in shifting standards (Biernat, Fuegen, & Kobrynowicz, 2010; Biernat & Kobrynowicz, 1997) or reconstructing criteria for evaluation (Uhlmann & Cohen, 2005). This approach builds on work illustrating that people behave more rationally when making joint vs. single evaluations (Bazerman, Loewenstein, & White, 1992; Hsee, Loewenstein, Blount, & Bazerman, 1999).

Placing applicants side-by-side could be an effective and easily implemented intervention to reduce bias in judgment. In Study 5, we investigated whether presenting applicants side-by-side reduced the criterion bias favoring more attractive people demonstrated in Studies 1a-1d.

Table 3
Correlations between Study 4 measures.

	Criterion Bias	Implicit Att.	Explicit Att.	Perceived Perf.	Desired Perf.
Implicit Att.	0.29				
Explicit Att.	0.41	0.31			
Perceived Perf.	0.62	0.24	0.49		
Desired Perf.	0.52	0.21	0.48	0.55	
Dating Att.	0.43	0.30	0.49	0.44	0.40

Note: Criterion Bias = criterion difference between White and non-White profiles. Implicit Att = aggregate MC-IAT *D* score for Whites. Explicit Att = aggregate explicit preference for Whites vs. Non-Whites. Perceived Perf = aggregate perceived performance for Whites vs. Non-Whites. Desired Perf = aggregate desired performance for Whites vs. Non-Whites. Dating Att. = single-item measure of interracial dating attitudes. All correlations significant at $p < .001$.

6.1. Methods

6.1.1. Participants

In exploratory analysis of Studies 1a–1c, we observed that criterion bias was strongest among younger participants. For Study 5, we limited eligibility to participants who were at most 30 years old. We sought to collect 2500 participants from the Project Implicit research pool, with 500 participants per experimental condition. Within each condition, this provided over 99% power for detecting a small (Cohen's $d = 0.20$) within-subjects effect size, and > 88% power at detecting a $d = 0.20$ effect between any two conditions.

Since studies on Project Implicit at the time were taken down on fixed days, the final sample was larger: 2855 participants volunteered, consented, and provided data. Among those who provided data, 65% were female, 65.0% were White, and the mean age was 22.2 ($SD = 3.92$). Sample sizes vary among tests due to missing data.

6.1.2. Procedure

The study session consisted of three components completed in the following order: the academic decision-making task, a survey about task performance and attractiveness preferences, and a measure of implicit attitudes towards more and less attractive people. See https://osf.io/rv6k5/ for the study's pre-registration.

6.1.2.1. Academic decision-making task. Participants completed an academic decision-making JBT measuring preferences for more or less physically attractive people, as in Study 1b. All participants first completed an encoding phase, where all 64 applicants were presented one at a time for 1 s each. Participants were then randomly assigned to one of five versions of the JBT. Within each condition, participants were assigned to complete one of eight orders. Across orders, each face was equally likely to be assigned to a more or less qualified application.

In the *Control* condition (64 trials; $n = 651$), participants completed the same single-evaluation JBT as Studies 1b–1d.

In the remaining four experimental conditions, applicants were shown in pairs (32 trials) with four response options: Accept both, Accept left, Accept right, and Reject both. In the *Just Comparison* ($n = 598$) condition, each pair consisted of two applicants that had the same level of qualification and attractiveness (e.g., two more attractive, more qualified applicants). In the *Cross-Attractiveness* condition ($n = 554$), each pair consisted of two applicants that had the same level of qualification and differing levels of attractiveness (e.g., two more qualified applicants, one more attractive and one less attractive). In the *Cross-Qualification* condition ($n = 507$), each pair consisted of two applicants that had the same level of attractiveness but differing levels of qualification (e.g., two more attractive applicants, one more qualified and one less qualified). Finally, in the *Fully Crossed* condition ($n = 545$), each pair consisted of two applicants that had differing levels of attractiveness and qualification (e.g., one more qualified, more attractive applicant and one less qualified, less attractive applicant).

Each of these experimental conditions had 32 total trials; 24 were critical trials described above and eight were distractor trials. Critical trials always consisted of faces from the same gender. Distractor trials always consisted of faces from different genders, so that the matching of genders in the critical trials was less obvious. Within each order, the applicants and images used as distractors were the same.

We did not analyze data from the distractor trials. In addition, to increase comparability between conditions, we did not analyze applicants in each order of the Control condition that were the distractor applicants in the experimental conditions from that same order, resulting in 48 critical trials in the *Control* condition. That is, within each order, all comparisons between *Control* and experimental conditions involved responses towards the same 48 face-applicant pairings. Results then compare decisions on the same applicants in each condition, with the only change being the context in which the applicants were judged.

6.1.2.2. Perception of performance and explicit preferences. Participants completed the same three items about perceived performance, desired performance and attractiveness preferences as in Study 1a.

6.1.2.3. Implicit preferences. Participants completed the same BIAT as in Study 1b.

6.2. Results

In Study 5, we first examined biases in criterion among all eligible participants, and whether the size of these criterion biases differed between experimental conditions. We then tested whether any experimental conditions differed in overall levels of sensitivity, explicit attitudes, implicit attitudes, perceived performance and desired performance. Finally, we analyzed how biases in criterion related to explicit attitudes, implicit attitudes, perceived performance and desired performance.

167 participants (5.9%) were excluded from analysis for accepting < 20% or > 80% of the applicants, or for accepting every more attractive or less attractive applicant. 80 additional participants (3.4% of those completing the BIAT) were excluded from analyses involving the BIAT for having > 10% of responses faster than 300 ms. The average acceptance rate was close 50% ($M = 50.7\%$, $SD = 12.0$). Participants required 7.25 min on average ($SD = 3.46$) to complete the task. The internal reliability of the criterion measure was comparable for more attractive ($\alpha = 0.59$) and less attractive ($\alpha = 0.61$) applicants. The reliability of the criterion difference score was $\alpha = 0.48$.

6.2.1. Criterion bias in decision-making

We tested whether there was evidence of differences in criterion between more and less attractive applicants within each condition. All conditions showed reliably lower criterion for more attractive relative to less attractive applicants, all t's > 5.59, all p's < 0.001, all d's > 0.24. There were no reliable differences in sensitivity between more and less attractive applicants, all t's < 1.38, all p's > 0.167, all d's < 0.06. See Table 4 for means, standard deviations and test statistics in each condition.

We next tested whether any experimental condition differed in their level of criterion bias relative to the *Control* condition. We again computed a criterion difference score (less attractive criterion – more attractive criterion), such that higher values meant lower criterion for more relative to less attractive applicants. When comparing criterion

Table 4
Sample sizes and criterion values in Study 5.

Condition	N	More attractive c	Less attractive c	d [95% CI]
Control	595	−0.07 (0.44)	0.07 (0.45)	0.29 [0.21, 0.38]
Just comparison	569	−0.13 (0.42)	0.07 (0.43)	0.43 [0.35, 0.52]
Cross-attractiveness	523	−0.11 (0.45)	0.02 (0.46)	0.24 [0.16, 0.33]
Cross-qualification	478	−0.10 (0.44)	0.05 (0.46)	0.34 [0.24, 0.43]
Fully crossed	523	−0.07 (0.46)	0.06 (0.49)	0.24 [0.16, 0.33]

Note: Criterion means and standard deviations within each condition of Study 5. All p values < .001.

bias across conditions, the only reliable result was a small effect such that participants in the *Just Comparison* condition showed slightly higher levels of criterion bias ($M = 0.20$, $SD = 0.46$) than participants in the

Control condition ($M = 0.14$, $SD = 0.49$), $t(1162) = 1.96$, $p = .050$, $d = 0.12$, 95% CI [0.0001, 0.23]. No other experimental conditions reliably differed from the *Control* condition, all t's < 0.36, all p's > 0.723, all d's < 0.02.

The *Just Comparison* condition also showed slightly higher levels of criterion bias than the *Cross-Attractiveness* condition ($M = 0.13$, $SD = 0.55$), $t(1090) = 2.07$, $p = .039$, $d = 0.13$, 95% CI [0.01, 0.24], and the *Fully Crossed* condition ($M = 0.14$, $SD = 0.54$), $t(1090) = 2.06$, $p = .040$, $d = 0.12$, 95% CI [0.01, 0.24], but did not reliably differ from the *Cross-Qualification* condition ($M = 0.15$, $SD = 0.46$), $t(1045) = 1.56$, $p = .120$, $d = 0.10$, 95% CI [−0.02, 0.22]. With multiple tests and weak effects, this suggests little meaningful variation in relative criterion bias across conditions.

Within each condition, sensitivity (d′) did not differ between more and less attractive applicants (all t's < 1.38, all p's > 0.167). However, we found large and intuitive differences between conditions on task sensitivity. Relative to the *Control* condition ($M = 0.98$, $SD = 0.53$), the *Just Comparison* condition showed lower sensitivity ($M = 0.71$, $SD = 0.52$), $t(1162) = 8.82$, $p < .001$, $d = 0.52$, 95% CI [0.40, 0.63], as did the *Cross-Attractiveness* condition, ($M = 0.66$, $SD = 0.55$), $t(1116) = 9.87$, $p < .001$, $d = 0.59$, 95% CI [0.47, 0.71]. Conversely, relative to the *Control* condition, the *Cross-Qualification* condition showed higher sensitivity, ($M = 1.25$, $SD = 0.68$), $t(1071) = 7.35$, $p < .001$, $d = 0.45$, 95% CI [0.33, 0.57], as did the *Fully-Crossed* condition, ($M = 1.30$, $SD = 0.59$), $t(1116) = 9.44$, $p < .001$, $d = 0.57$, 95% CI [0.45, 0.69]. In the latter two conditions, the side-by-side profiles differed in qualification, making those differences easier to detect. In the former two experimental conditions, the side-by-side profiles had the same overall qualification, making it more difficult to accurately detect qualification differences across trials.

6.2.2. Differences in attitudes, desired and perceived performance

Explicit attitudes indicated preference for more attractive people ($M = 0.92$, $SD = 1.01$, $d = 0.91$). Relative to the *Control* condition, there were no reliable differences in explicit attitudes across conditions, all t's < 0.53, all p's > 0.597, all d's < 0.03. Implicit attitudes indicated more positive associations towards more attractive people ($M = 0.70$, $SD = 0.48$, $d = 1.46$). Relative to the *Control* condition ($M = 0.74$, $SD = 0.46$, $d = 1.61$), participants in the *Fully Crossed* condition ($M = 0.68$, $SD = 0.50$, $d = 1.36$) showed slightly lower levels of implicit positive associations towards more attractive people, $t(931) = 2.05$, $p = .041$, $d = 0.13$, 95% CI [0.01, 0.26]. No other experimental conditions differed from the *Control* condition, all t's < 1.52, all p's > 0.129, all d's < 0.10. The single positive result seems likely to be a false positive.

We next tested for whether any conditions differed from the *Control* condition in the proportion reporting having shown no bias on the task. Relative to the *Control* condition (66.7%), participants in the *Cross-Attractiveness* condition (72.7%) showed a small increase in proportion reporting having treated all applicants equally, $\chi^2(1, N = 1079) = 4.60$, $p = .034$, as did participants in the *Cross-Qualification* condition, $\chi^2(1, N = 1033) = 4.18$, $p = .042$. Neither the *Just Comparison* (72.7%) nor the *Fully Crossed* condition (71.7%) differed from the *Control* condition in proportion reporting a perception of being fair, all $\chi^2 > 3.10$, p's > 0.085.

Finally, we tested whether any experimental conditions differed from the *Control* condition in proportion reporting a desire to show no bias on the task. Relative to the *Control* condition (82.8%), participants in the *Cross-Qualification* condition (87.7%) showed a small increase in proportion of reporting wanting to treat all applicants equally, $\chi^2(1, N = 1038) = 4.94$, $p = .029$. Neither the *Just Comparison* (85.1%), the *Cross-Attractiveness* condition (86.3%) or the *Fully Crossed* condition (85.7%) differed from the *Control* condition in proportion reporting a perception of being fair, all $\chi^2 > 2.54$, p's > 0.11.

6.2.3. Predicting criterion bias

We computed the same criterion difference score as in Studies 1a-1d. Across conditions, this difference score was reliably correlated with BIAT D scores ($r(2274) = 0.13$, $p < .001$, 95% C.I. [0.08, 0.19]), explicit preferences for more attractive people ($r(2600) = 0.13$, $p < .001$, 95% C.I. [0.08, 0.18]), perceptions of performance ($r(2595) = 0.27$, $p < .001$, 95% C.I. [0.23, 0.32]), and desired performance ($r(2600) = 0.14$, $p < .001$, 95% C.I. [0.10, 0.17]).

A simultaneous linear regression with implicit attitudes, explicit attitudes and condition (coded with *Control* as the reference) predicting criterion bias revealed that implicit ($b = 0.12$, $t(2245) = 5.56$, $p < .001$) and explicit attitudes ($b = 0.05$, $t(2245) = 4.97$, $p < .001$) reliably predicted differences in response criterion, but condition did not (all b's < 0.05, all t's < 1.86, all p's > 0.063; see online supplement for full reporting). Another simultaneous linear regression adding perceived and desired performance revealed that implicit attitudes ($b = 0.10$, $t(2232) = 4.52$, $p < .001$), explicit attitudes ($b = 0.02$, $t(2232) = 2.34$, $p = .020$), perceived performance ($b = 0.14$, $t(2232) = 11.02$, $p < .001$), and desired performance ($b = 0.07$, $t(2232) = 3.89$, $p < .001$) contributed uniquely, while experimental condition did not (all b's < 0.05, all t's < 1.86, all p's > 0.063). These variables accounted for 9.5% of the difference in criterion bias.

6.3. Discussion

Participants had a lower criterion for more than less attractive applicants, and this did not differ between conditions of single or joint-evaluation. One condition (*Just Comparison*) showed a small *increase* in criterion bias compared to *Control*. Moreover, two joint evaluation conditions (*Cross-Attractiveness* and *Cross-Qualification*) showed slightly higher rates of perceiving treating more and less attractive applicants equally than the *Control* condition, despite showing comparable levels of criterion bias (e.g., Lindner, Graser, & Nosek, 2014; Norton et al., 2004).

Joint evaluation impacted sensitivity (i.e., accuracy). For conditions in which applicants in each comparison were equally qualified, participants were significantly *less* accurate than the *Control* condition. Conversely, in conditions where applicants in each comparison were differentially qualified, participants were significantly *more* accurate than the *Control* condition. Notably, higher or lower accuracy did not substantially alter criterion bias.

Study 5 results suggest some durability of the biases measured by the JBT. Even when more and less qualified applicants were presented side-by-side, participants were still more lenient towards more attractive applicants. This occurred despite most participants reporting a desire to be fair (86.7%) and a perception of having been fair (72.1%). These data might suggest that joint evaluation is less effective when the potential bias is not highly accessible or obvious to participants. That is, participants may be less likely to spontaneously recognize physical attractiveness as a potentially biasing influence compared to gender or race (i.e., the social dimensions most frequently studied). Another possibility is that the benefits of joint evaluation are less effective when participants make many judgments. These speculations require direct investigation to assess their viability.

7. General discussion

Social biases in judgment become problematic when they differ from conscious values and can occur without the intent to discriminate or the awareness of having done so (Bertrand et al., 2005). The existence and operation of intended and unintended biases is subject to intense research efforts, but common measures using single judgments, lacking an objective standard, and being inflexible significantly limit

the pace of knowledge accumulation. To improve measurement of social judgment biases, we developed the Judgment Bias Task (JBT). The JBT has a flexible structure for a variety of uses, contains multiple judgments, has an objective standard to identify whether bias had occurred, assesses individual differences in the magnitude of bias, and takes an average of 6 min.

Using Signal Detection Theory, the JBT identified social judgment biases of greater leniency (lower criterion) for honor society candidates that were more attractive than less attractive (Studies 1a–1d, 5), from one's own than another university (Studies 2a & 2b), and from one's own than another political party (Study 3). In a dating context, White participants had a lower criterion for dating members of their own race compared to dating Blacks and Hispanics (Study 4). Criterion biases persisted even when more and less qualified applicants were presented side-by-side (Study 5).

Criterion biases were often present among participants who reported having no explicit preference, not showing favoritism, or not wanting to show favoritism. This suggests that the expressed biases sometimes occurred without intention or awareness. Simultaneously, criterion bias was correlated with perceived performance, desired performance, and explicit attitudes, suggesting that the expressed biases were related to intention and awareness. Together, social judgment biases on the JBT may be influenced both by intentional and unintentional mental processes and are partially though not completely available to awareness and control.

7.1. Using the JBT to advance theory and evidence about social judgment biases

Efficient, effective measurement methods can accelerate theoretical progress by providing a flexible, repeatable investigation framework. The JBT was sensitive to measuring well-known social biases, revealing selection preferences for more physically attractive people (Beehr & Gilmore, 1982; Studies 1a-1d and 5) and for one's ingroup (Mullen et al., 1992; Studies 2a, 2b and 3), and for members of one's own race in a romantic context (Gaines Jr., Gurung, Lin, & Pouli, 2006; Study 4). The JBT can be adapted for measuring social judgment biases about other groups such as age, gender, or religion, and about selection for other social, performance, or leadership outcomes. For example, in technology-based professions, there is documented bias for hiring younger applicants (McCann & Giles, 2002). To measure possible individual differences in this age-based bias, the JBT could be adapted to screen new employees for a technological company and then alter applicants ages and tech-related qualifications.

The JBT can also be adapted to experimentally investigate contextual or procedural factors in social judgment biases. In Study 5, for example, we examined whether social biases changed as a function of single versus joint evaluation. Other potential manipulations include (1) varying the proportion of candidates from social groups to examine minority/majority effects, (2) varying the quality of candidates between social groups such as a design in which Black college applicants have weaker academic credentials on average than White college applicants, (3) including "distractor" profiles of extremely qualified or not qualified candidates to investigate anchoring of social judgments, and (4) comparing the magnitude of bias observed between-subjects versus within-subjects assessments (e.g., comparing criterion across conditions that only viewed more or less physically attractive applicants). For the latter, within-subjects comparisons might have produced contrast effects that could exacerbate biases (e.g., Hosoda et al., 2003).

Performance on the JBT may also be affected by judgment context. For example, time provided for decision-making might influence the likelihood that judgment biases are expressed. In a meta-analysis across studies, there was a small but reliable association between average (log-transformed) time spent on each judgment and overall criterion bias (aggregate $r = -0.07$, 95% CI $[-0.11, -0.04]$, see online supplement for full details). This weak effect is consistent with the hypothesis that biases that are difficult to control may increase under time pressure. This result is a post hoc observation and correlational, but the JBT could be adapted to conduct an experimental test.

Furthermore, the JBT could be used to test the effectiveness of various bias reduction interventions. The replicable effects found here (average criterion $d = 0.33$) are well suited to investigate the relative strength of various biased behavior reduction strategies, similar to Lai et al.'s (2014) "contest" for testing interventions to reduce implicit racial biases. Having a replicable measure of social judgment biases could rapidly accelerate theoretical and empirical advances. Simple changes and interventions to the JBT procedure can provide efficient experimental methods for advancing understanding the impact of bias reduction strategies such as creating a common ingroup (Gaertner, Mann, Murrell, & Dovidio, 1989), using implementation intentions (Mendoza, Gollwitzer, & Amodio, 2010), or increasing accountability (Webster, Richter, & Kruglanski, 1996).

Experimental manipulations of the JBT may help identify the contexts, individual differences, and mechanisms that shape expression of bias via intentional or unintentional processes. Participants can easily choose to express an explicit bias in the JBT by selecting candidates of one social category and rejecting candidates of another. However, participants may not easily choose to not express an implicit bias in the JBT because they do not recognize its operation or know how to correct it. The experimental control afforded by the JBT offers opportunity to systematically evaluate the processes that promote and mitigate operation of implicit biases on ostensibly explicit judgments like selection, hiring, or voting.

Finally, because the JBT has an objective accuracy standard, it will facilitate investigation of whether such strategies are *debiasing*, or actually reverse bias to favor another group. In short, by providing a replicable measure of biased behavior, the JBT offers an efficient means to refine, refute, and generate theoretical knowledge about social bias in behavior (Greenwald, 2012).

7.2. JBT as a predictor of other social biases?

In this paper, the JBT was used exclusively as an outcome measure. We examined the effect of experimental interventions on the expression of bias on the JBT, and the ability of other variables to predict variation in JBT performance. Also, the research applications that we propose above treat the JBT as an outcome measure for investigating theoretical interests in the processes underlying social judgment biases.

It is conceivable that the JBT could be used productively as an independent variable to predict other forms of social judgment and behavioral biases. Next steps for such research applications would be to further clarify the convergent and discriminant validity of the JBT with other measures of social bias. Also, we do not yet have evidence concerning the JBT's external test-retest reliability or stability over time. Such evidence would be useful for understanding the JBT's potential as an individual differences predictor of social biases that occur across time.

7.3. Is the JBT an implicit measure?

No. Respondents on the JBT can control and directly express their social biases. For example, if participants explicitly wanted to exclude students from a rival school from an academic honor society, they could apply that desire easily when performing the JBT. In other words, it is straightforward for JBT respondents to apply an explicit decision rule about a social group just as they could for any real-world hiring decision or related judgment.

The JBT is not an implicit measure, but that does not address whether performance on the JBT can be influenced by implicit processes. Participants who wanted to avoid favoring one group over another can fail to do so. Even among participants that did not want to be biased or believe that they behaved in an unbiased manner, the JBT

consistently revealed social biases in judgment.

It is important to understand this asymmetry. To qualify as an implicit measure, assessment of the association of interest must be indirect (Greenwald & Banaji, 1995; Nosek & Greenwald, 2009). But, any behavior could be influenced by implicit processes. For example, people may be unknowingly influenced by race (McDermott, 1998), height (Sorokowski, 2010), or attractiveness (Banducci, Karp, Thrasher, & Rallings, 2008) in voting behavior. That does not make voting an implicit measure. The present evidence suggests that the JBT is sometimes influenced by biases occurring outside of the respondents' awareness or control. A key strength of the JBT is the opportunity to use the paradigm to investigate the conditions under which implicit processes will be more or less influential on social judgment. This may provide an experimental testbed for developing theories for how these processes operate in consequential domains like hiring, voting, and other selection decisions.

7.4. Methodological strengths and limitations of JBT for investigating social judgment biases

The accumulated evidence highlights several methodological strengths and limitations of the JBT. First, the JBT is flexible for assessing biases about different social groups and a variety of outcomes. The JBT is also highly adaptable for investigating procedural factors, such as the number of comparison groups or type of judgment made. For example, in these studies, participants made binary accept/reject judgments, but one could test whether biases increase or decrease if participants have more response options for expressing judgment, such as a Likert scale. We adopted a binary response to maximize the benefits of Signal Detection Theory for modeling criterion and sensitivity. Changes to procedural features of the JBT may require consideration of alternative measurement and analytic strategies modeling the sources of social bias. It is unknown whether such changes would strengthen or weaken the JBT's sensitivity to assessing judgment biases.

Second, by using a within-subjects design with multiple trials, the JBT is sensitive to individual differences and can estimate the magnitude of bias compared to an objective standard indicating no bias. Using SDT, a criterion value of zero towards one social group indicates equal likelihood of correctly accepting more qualified profiles and correctly rejecting less qualified profiles from that group, and no difference in criterion values between two groups indicates that a participant applied the same degree of leniency to profiles from both groups. A zero for the criterion difference score provides an unambiguous interpretation that criterion levels did not differ between social groups, meaning there was no evidence of bias. Of course, this does not suggest that biases towards these groups do not exist or would not occur in other contexts. Of interest is how JBT values relate to other measures of performance, such as attitude measures or participants' own perceived and desired performance.

Effective assessment with the JBT requires careful attention to the design and characteristics of the profiles. We selected dimensions that were similarly useful so that participants would find it reasonable to weigh them equally. This is important as the objective standard for determining accuracy requires adherence to the equal weighting instruction. We also selected stimuli and varied scaling so that participants would find it relatively difficult to distinguish between more and less qualified applicants (median 67% accuracy across studies). This is important because social biases may be weaker when differences between more and less qualified profiles are easy to detect (e.g., Dovidio & Gaertner, 2000). We suspect that effective use of the JBT will require close attention to design of these features for each application. The supplementary materials provide substantial detail to facilitate additional use.

Third, the JBT was efficient to administer. Across all studies, the median average completion time, including instructions, was 5.80 min. This is short enough for many research applications, but may still be a barrier for data collection with expensive samples. It is conceivable that study administration can be shortened further. However, a cost could be reliability of measurement with fewer trials. Here, we used a minimum of 60 total trials and 20 trials per group. More generally, it would be useful to investigate the number of trials in the JBT to optimize reliability, validity, and time of administration. The present data could provide a starting point by conducting trial-level analyses with the JBT (data available at https://osf.io/u2mbx/). Other methodological features that could be examined include exclusion criteria to simultaneously maximize participant retention and minimize the inclusion of inattentive participants.[8]

To facilitate wider use of the JBT, we have developed an Inquisit program for data collection and syntax in SPSS, R and SAS for data analysis with SDT. We have also written a “how to” guide with step-by-step instructions for creating new versions of the JBT. These materials are available at jordanaxt.com and https://osf.io/u2mbx/.

7.5. Reliability of the JBT

The internal reliability of the JBT towards any one social group was moderately high (median $\alpha = 0.61$), and the internal reliability of the criterion difference score was moderate but weaker (median $\alpha = 0.48$, minimum $\alpha = 0.14$, maximum $\alpha = 0.78$). The literature identifies challenges for interpreting the reliability of difference scores, with some arguing against their use (e.g., Cronbach & Furby, 1970; Peter, Churchill Jr, & Brown, 1993). More recent work has defended the use of difference scores, given that some of the assumptions applied to other measures do not hold for difference scores. For instance, reliability for difference scores decreases as the correlation between the component scores increases (Thomas & Zumbo, 2012). Likewise, greater similarity in variances between component scores will also decrease the reliability of the difference score (Trafimow, 2015). Williams and Zimmerman (1996) argue that these features lead to underestimating the reliability of difference scores compared to other measures. For example, in Study 2a, the JBT showed substantially stronger correlations with some of the predictor variables than the estimated internal reliability of the measure.

Nevertheless, there may be opportunities to increase the reliability of JBT with procedural innovations–an obvious one being adding trials to the procedure. Also, increasing the encoding time to give participants a better understanding of the range of qualifications or other procedural innovations may likewise increase the JBT's reliability. However, each effort to increase reliability will need to weigh against other qualities of design and construct validity. Certain elements of the JBT that make the task more effective for testing social bias may necessarily reduce its reliability. For one, the task is designed to be difficult, comparing applicants that have comparable objective qualifications. This ambiguity makes it more likely that irrelevant social factors may impact judgment (e.g., Dovidio & Gaertner, 2000), but also makes responses less consistent and reliable because, by design, they are more likely to be influenced by extraneous factors. If we had compared highly unqualified and highly qualified applicants, accuracy and reliability would likely improve, but by eliminating ambiguity and the potential influence of irrelevant social information.

In addition, all profiles were unique and scored to have the same level of either high or low qualifications. This makes variation across profiles more realistic for the participant by deliberately introducing noise. Profiles with the same overall qualification score can vary in accuracy because of variation in difficulty (e.g., profiles with round numbers may be more appealing regardless of qualification strength). If all of the more and less qualified profiles had the same academic values,

[8] These studies used the same JBT exclusion criteria as earlier work (e.g., Axt et al., 2016). However, the online supplement reports analyses of criterion differences using all participants, which did not substantively alter any conclusions.

reliability would again increase at the cost of reducing ambiguity. In short, there may be an effective ceiling on reliability when judgment biases depend, in part, on answers not being obvious.

Finally, the influence of irrelevant social biases on judgment may "naturally" be a relatively low reliability behavior. Most respondents try to avoid use of irrelevant social information most of the time for making social judgments. If the influence of irrelevant information occurs intermittently, particularly uncontrollably, then reliability will necessarily be relatively low because participants are successfully using the objective criteria most of the time. As such, for realistic investigations of how social biases operate, researchers may need to assume and prepare for the possibility that the outcome of interest will occur intermittently and power their studies accordingly.

In summary, optimizing the JBT's design features might increase reliability and sensitivity to judgment biases. That said, even with moderate internal reliability, we observed relatively large mean-level effects of criterion bias ($d = 0.57$ in Study 1a) and moderate correlations with explicit attitudes ($r = 0.31$ in Study 3) and perceptions of performance ($r = 0.43$ for desired and $r = 0.46$ for perceived performance in Study 4). This suggests that the JBT is already an efficient measure for conducting relatively high-powered research on social judgment biases.

7.6. Methodological analyses of the JBT

There are a variety of methodological and procedural elements of the JBT that may be important for effective design and measurement. We examined some by conducting exploratory analysis using data from Study 1b. After identifying the most interesting methodological issues or features, we replicated those analyses on all other studies in which participants evaluated applicants one at a time (Studies 1a– 4). We provide a brief summary of analyses here, and full details are available in the online supplement.

7.6.1. Exclusion criteria

For all studies, our pre-registered analysis plan excluded participants who accepted < 20% or > 80% of the profiles, based on thinking those cutoffs would exclude inattentive participants who deviated too far from the suggested 50% acceptance rate. We tested whether results differed based on differing exclusion criteria, and found little variation. For example, in Study 3, comparing all participants versus only those meeting the 20%–80% acceptance rate cutoff found little differences in overall accuracy (All = 66.7%; 20%–80% = 67.9%), size of the ingroup criterion bias (All $d = 0.32$; 20%–80% $d = 0.31$), or correlation with explicit attitudes (All $r = 0.33$; 20%–80% $r = 0.32$) or implicit attitudes (All $r = 0.21$; 20%–80% $r = 0.22$; see online supplement for analyses from all studies using a variety of exclusion criteria). These results suggest future analyses of the JBT may focus on finding suitable exclusion criteria (e.g., using average reaction time per judgment), and this work may improve the JBT's reliability. The datasets from these studies are available to help initiate such an investigation.

7.6.2. Accuracy by trial number

Across studies, there was variation in the degree to which accuracy in accepting more qualified and rejecting less qualified applicants increased or decreased over time. For example, in Study 1b, mean accuracy was 66.6% (Range 64.6%–69.8%) and accuracy was negatively correlated $r(64) = -0.43$ with trial number. In Study 1d, mean accuracy was 64.2% (Range 61.2%–67.4%) and accuracy was slightly positively correlated $r(64) = 0.11$ with trial number. Aggregating across studies, there was no consistent relationship between trial number and accuracy, $r = -0.06$, 95% CI [−0.15, 0.03]. Given the cumulative evidence, we presently do not believe that accuracy improves with 60 or fewer trials of experience.

7.6.3. Criterion bias across the task

Criterion bias remained steady between the first and second half of the JBT. We divided each participant's responses into two sets depending on when each applicant/social group combination was encountered and calculated criterion for each set. For all studies, criterion biases emerged in both the first and second sets, and in general, the strength of the criterion bias did not reliably differ between sets. For example, in Study 1c, participants showed a lower criterion for more than less attractive applicants in both the first set ($t(874) = 7.92$, $p < .001$, $d = 0.28$) and the second set ($t(874) = 7.46$, $p < .001$,$d = 0.24$), and these did not differ ($t(874) = 0.46$, $p = .648$, $d = 0.02$). Given the cumulative evidence, we presently do not believe that the magnitude of judgment biases changes over the course of 60 or fewer trials.

7.6.4. Criterion bias by face-applicant pairings

Several of the studies (1b–d, 2 & 4) randomly assigned participants to one of 12 or 18 previously created pairings between faces (or social groups) and profiles. In each study, half of the pairings were randomly generated and the other half generated by switching the social groups assigned to each applicant. Across studies that randomly assigned participants to previously created pairings, there was little evidence for a main effect of pairing on criterion; that is, pairings did not alter the overall criterion. However, in each study, there was evidence for an influence of pairing on the difference in criterion between social groups (average $\eta_p^2 = 0.06$). This suggests that some combinations of applicants and social groups may have elicited stronger effects than others. Our reported aggregate effects were robust to pairing effects, and studies that did not use previously created pairings (Studies 1a, 2a, and 2b) also found biases in criterion. However, these analyses highlight the importance of randomization, and suggests investigating the influence of the criteria combinations on accept/reject judgments to reduce pairing effects.

7.6.5. Variation in profile accuracy

One possibility for producing order effects is if some profiles were easier to evaluate than others, eliciting systematically smaller criterion bias effects. We did observe evidence for variation in profile accuracy (e.g., $M = 64.2\%$, $SD = 13.8$, Range = 40.2% to 85.3% in Study 1d), with accuracy on some profiles even being below chance. Notably, criterion bias effects were observed for high and low accuracy profiles. In Study 1b, the differences in accuracy for each profile when paired with more vs. less attractive faces were consistent with the biases in criterion. For every less qualified applicant, accuracy was higher when paired with less attractive than more attractive faces (meaning applicants were more likely to be incorrectly admitted when paired with a more attractive face). Conversely, for every more qualified applicant, accuracy was higher for more attractive than less attractive faces (meaning applicants were more likely to be incorrectly rejected when paired with a less attractive face).

Similarly, the reliability of the criterion difference score did not change significantly when excluding profiles that were either low or high in accuracy. For example, in Study 3, the reliability found when using all profiles ($\alpha = 0.61$) was if anything higher than the reliability calculated after excluding profiles with lower than 50% accuracy ($\alpha = 0.59$) or lower than 55% accuracy ($\alpha = 0.55$; see online supplement for similar analyses for all studies). These results suggest that calibration of profile difficulty may enhance criterion bias effects but is not essential for observing or measuring them.

7.6.6. Variation in face accuracy

There was less evidence for systematic variation in accuracy by faces in Studies 1a-1d (e.g., $M = 70.4\%$; $SD = 3.3$; *Range* = 62.3% to 77.5% in Study 1a), suggesting that meaningful variability in accuracy is more a function of the applicant qualifications than the faces used.

7.6.7. Influence of criteria on accept/reject decisions

For studies 1a-4, we ran Hierarchical Linear Models (HLM), with trials nested within participants, predicting the likelihood of an accept decision from each of the listed criteria (uncentered), placing all criteria on a 1–4 scale. In each study, all criteria independently predicted acceptance decisions (all t's > 14.24), such that higher scores on each were associated with a greater likelihood of the applicant being accepted. However, we found evidence that some criteria were more influential than others. For example, in Study 3, science GPA ($b = 2.27$) and interview score ($b = 3.07$) had a stronger influence on accepting the applicants than did humanities GPA ($b = 1.38$) and recommendation letters ($b = 0.77$). This appears to account for variation in profile accuracy, and could be the source of differences in criterion bias by task pairings.

These results emphasize the importance of counterbalanced designs - ideally within and between subjects - such that participants from each group are equally represented with better and worse scores on each criterion across the profiles. Also, these results suggest opportunities to improve profile criteria to increase the consistency of their influence on decision-making.

7.6.8. Importance of encoding

All studies included an encoding phase where participants were briefly shown all applicants or profiles before making any accept or reject decisions. To test the necessity of this encoding phase, we ran a tenth study in which online participants ($N = 801$) completed Study 1b measures either with or without the encoding phase. See https://osf.io/eg6f9/ for the study's pre-registration and the online supplement for full methods and results. In both conditions, participants had lower criterion for more versus less physically attractive applicants (Encoding: $t(412) = 5.70$, $p < .001$, $d = 0.28$, No-Encoding: $t(388) = 5.87$, $p < .001$, $d = 0.30$), and no reliable difference in the size of the criterion bias between conditions ($t(799) = 0.04$, $p = .971$, $d = 0.003$). There was also no evidence that the two conditions differed on task sensitivity (Encoding: $M = 0.96$, $SD = 0.61$; No-Encoding: $M = 0.96$, $SD = 0.60$; $t(799) = 0.04$, $p = .972$, $d = 0.002$).

These findings suggest the encoding phase may have little impact on the degree of social bias or ability to distinguish between more and less qualified applicants. Removing the encoding phase shortens the average time to complete the JBT by 95 s and, based on this study, had little deleterious effect on measurement quality. However, removing encoding also increased the percentage of excluded participants (21.9% without encoding vs. 8.6% with encoding). This suggests that a shorter task comes with a tradeoff of, at minimum, losing power because of increased participant exclusion. There may be alternative strategies to be discovered that provide instructions more rapidly without the deleterious impact on participant exclusion rates.

7.6.9. Alternative analysis strategies

We used SDT for analysis because of the benefits for separating sensitivity (ability to detect more from less qualified) and criterion (likelihood of selecting someone as qualified). However, this is not the only choice for analyzing these data. Trial-level analysis using HLM could predict the likelihood of an "accept" decision based on applicant qualifications and social group, leveraging some of the reliability and sensitivity benefits of this analytic strategy (Hox, 1998). The coefficient for the social group variable would be similar to the criterion bias analysis reported here.

For each study, we ran such HLM analyses among 1) all eligible participants, 2) participants reporting a desire to be fair, and 3) participants reporting a perception of having been fair. In all studies, conclusions from HLM analyses mirrored those using SDT. The one exception was Study 4, in which an HLM analysis among participants who perceived treating all races equally now showed a small but reliable ingroup bias in evaluation ($b = 0.012$, $t(576) = 2.41$, $p = .016$); in our SDT approach, this effect was not reliable ($t(576) = 1.57$, $p = .118$). While nicer for consistency with other observed ingroup effects, we do not perceive this difference as justifying deviation from our pre-registered SDT analysis plan for primary result reporting. All HLM analyses are available in the online supplement.

Another analysis strategy is to leverage the findings that reliabilities were generally higher for criterion towards the individual social groups in each study (e.g., more and less physically attractive applicants) than the reliability of the criterion difference score using the two social groups. For instance, instead of correlating the difference score with outcome variables like attitudes or perceived or desired performance, researchers could treat the two criteria as a repeated measure (i.e., a within-subjects factor of "social group") and run a mixed model to see if the random slope for the social group factor was moderated by any of these outcome variables.

We re-ran our analyses from Studies 1a–5 and compared results from correlating the difference score with the attitudes and performance measures with this mixed-model approach. In each of the 36 analyses, conclusions were the same in terms of rejecting the null hypothesis at $p < .05$ (32 analyses rejected the null, four analyses failed to reject the null; see online supplement for full results). We also compared effect sizes of the two analysis strategies, converting each into a Cohen's d. Across the 36 analyses, the two approaches produced nearly identical results, with the largest difference between the two being $d = 0.013$. Detail of these analyses is available at https://osf.io/zbksa/. It is possible that future investigations will reveal additional value from representing JBT data in a multi-level model. For the present studies, at least, we observed no gain compared to our simple difference score analysis strategy.

8. Conclusion

Social judgment biases are prevalent and often unintended. We introduced a research paradigm, the JBT, that revealed replicable biases in decision-making that sometimes occurred outside of conscious awareness or intention. The JBT is flexible and capable of examining individual differences in the magnitude of bias. The JBT builds on methodologies with features that could improve research efficiency investigating judgment biases but have not gained wide adoption (Beckett & Park, 1995; Caruso et al., 2009; Locksley et al., 1982). The results and resources presented here may lower the barrier to adoption. The JBT provides an efficient means of pursuing theoretical advances in assessing social judgment biases, how they are formed, and perhaps how they can be changed.

Open practices

This article earned the Preregistration, Open Materials, and Open Data badges for transparent research practices. Links to each pre-registration can be viewed at https://osf.io/j9qwy/. Materials and Data for all experiments are available at https://osf.io/u2mbx/.

Acknowledgments

We thank Caroline Hamby and Kaylee Nichols for assistance in data collection for Studies 1a and 2a, and Chelsea Schein for assistance in data collection for Study 2b.

Declaration of conflicting interests

This research was partly supported by Project Implicit. B. A. Nosek is an officer and both he and J. R. Axt are consultants of Project Implicit, Inc., a nonprofit organization with the mission to "develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender, or other factors." The authors declared that they had no other potential conflicts of interest with respect to their

authorship or the publication of this article.

Author contributions

All authors developed the concept of Studies 1a and 2a. J. R. Axt and H. N. Nguyen programmed Studies 1a and 2a and analyzed the data. For all other studies, J. R. Axt and B.A. Nosek developed the concept, with J.R. Axt programming them and analyzing the data. J. R. Axt drafted the manuscript, and B. A. Nosek and H.N. Nguyen edited it. All authors approved the final version of the manuscript for submission.

Author note

Portions of these results were presented at the 2016 conference for Society for Personality and Social Psychology. The data, materials, and manuscript have also been posted on the first author's website and on the OSF (https://osf.io/u2mbx/).

Chronology of studies

Studies are numbered 1–5 for narrative style. Chronologically, studies were run in the following order: 1a, 2a, 1b, 3, 4, 5, 2b, 1c, 1d.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2018.02.011.

References

Ameri, M., Schur, L., Adya, M., Bentley, S., McKay, P., & Kruse, D.. The disability employment puzzle: A field experiment on employer hiring behavior (working paper no. w21560). (2015). Retrieved from http://www.nber.org/papers/w21560 (National Bureau of Economic Research).

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108–119.

Axt, J. R. (2017). An unintentional pro-Black bias in judgment among educators. *British Journal of Educational Psychology*. http://dx.doi.org/10.1111/bjep.12156 (Advance online publication).

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science, 25*(9), 1804–1815.

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.

Banducci, S. A., Karp, J. A., Thrasher, M., & Rallings, C. (2008). Ballot photographs as cues in low-information elections. *Political Psychology, 29*(6), 903–917.

Bazerman, M. H., Loewenstein, G. F., & White, S. B. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly, 37*(2), 220–240.

Beckett, N. E., & Park, B. (1995). Use of category versus individuating information: Making base rates salient. *Personality and Social Psychology Bulletin, 21*(1), 21–31.

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*, 94–98.

Bertrand, M., & Duflo, E.. Field experiments on discrimination (working paper no. 22014). (2016). Retrieved from http://www.nber.org/papers/w22014 (National Bureau of Economic Research).

Biernat, M., Fuegen, K., & Kobrynowicz, D. (2010). Shifting standards and the inference of incompetence: Effects of formal and informal evaluation tools. *Personality and Social Psychology Bulletin, 36*(7), 855–868.

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

Binning, K. R., Brick, C., Cohen, G. L., & Sherman, D. K. (2015). Going along versus getting it right: The role of self-integrity in political conformity. *Journal of Experimental Social Psychology, 56*, 73–88.

Blommaert, L., van Tubergen, F., & Coenders, M. (2012). Implicit and explicit interethnic attitudes and ethnic discrimination in hiring. *Social Science Research, 41*(1), 61–73.

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

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376.

Cabrera, M. A. M., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*(1–2), 103–113.

Caruso, E. M., Rahnev, D. A., & Banaji, M. R. (2009). Using conjoint analysis to detect discrimination: Revealing covert preferences from overt choices. *Social Cognition, 27*(1), 128–137.

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.

Cheung, B. Y., & Heine, S. J. (2015). The double-edged sword of genetic accounts of criminality causal attributions from genetic ascriptions affect legal decision making. *Personality and Social Psychology Bulletin, 41*(12), 1723–1738.

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.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology, 92*(6), 1006–1023.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68–80.

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

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science, 11*(4), 315–319.

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

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

Forscher, P. S., Cox, W. T., Graetz, N., & Devine, P. G. (2015). The motivation to express prejudice. *Journal of Personality and Social Psychology, 109*(5), 791–812.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology. *Personality and Social Psychology Review, 18*(1), 3–12.

Gaertner, S. L., Mann, J., Murrell, A., & Dovidio, J. F. (1989). Reducing intergroup bias: The benefits of recategorization. *Journal of Personality and Social Psychology, 52*, 239–249.

Gaines, S. O., Jr., Gurung, A. R., Lin, Y.-Y., & Pouli, N. (2006). Interethnic relationships. In P. Noller, & J. Feeney (Eds.). *Close relationships: Functions, forms and processes* (pp. 171–186). New York: Psychology Press.

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science, 7*(2), 99–108.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4–27.

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

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

Haddock, G., Zanna, M. P., & Esses, V. M. (1993). Assessing the structure of prejudicial attitudes: The case of attitudes toward homosexuals. *Journal of Personality and Social Psychology, 65*, 1105–1118.

Hawkins, C. B., & Nosek, B. A. (2012). Motivated independence? Implicit party identity predicts political judgments among self-proclaimed independents. *Personality and Social Psychology Bulletin, 38*(11), 1437–1452.

Hosoda, M., Stone-Romero, E. 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.

Hox, J. (1998). Multilevel modeling: When and why. *Classification, data analysis, and data highways* (pp. 147–154). Springer Berlin Heidelberg.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576.

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*, 301–318.

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

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: IV: Procedures and validity. In B. Wittenbrink, & N. Schwarz (Eds.). *Implicit measures of attitudes: Procedures and controversies* (pp. 59–102). New York: Guilford Press.

Lindner, N. M., Graser, A., & Nosek, B. A. (2014). Age-based hiring discrimination as a function of equity norms and self-perceived objectivity. *PLoS One, 9*(1), e84752.

Locksley, A., Hepburn, C., & Ortiz, V. (1982). Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology, 18*(1), 23–42.

McCann, R., & Giles, H. (2002). Ageism in the workplace: A communication perspective. *Ageism: Stereotyping and prejudice against older persons* (pp. 163–200). Cambridge, Massachusetts: MIT Press.

McDermott, M. L. (1998). Race and gender cues in low-information elections. *Political Research Quarterly, 51*(4), 895–918.

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*, 512–523.

Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination: An audit study in academia. *Psychological Science, 23*, 710–717.
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.
Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology, 87*, 817–831.
Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go association task. *Social Cognition, 19*(6), 625–666.
Nosek, B. A., Bar-Anan, Y., Sriram, N., Axt, J., & Greenwald, A. G. (2014). Understanding and using the brief Implicit Association Test: Recommended scoring procedures. *PLoS One, 9*(12), e110938.
Nosek, B. A., Ebersole, C. R., DeHaven, A., & Mellor, D. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences* (in press).
Nosek, B. A., & Greenwald, A. G. (2009). (Part of) the case for a pragmatic approach to validity: Comment on De Houwer, Teige-mocigemba, Spruyt, and Moors (2009). *Psychological Bulletin, 135*(3), 373–376.
Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*(2), 166–180.
Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. *Automatic processes in social thinking and behavior* (pp. 265–292). New York: Psychology Press.
Peter, J. P., Churchill, G. A., Jr., & Brown, T. J. (1993). Caution in the use of difference scores in consumer research. *Journal of Consumer Research, 19*(4), 655–662.
Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17*(3), 523–534.
Sorokowski, P. (2010). Politicians' estimated height as an indicator of their popularity. *European Journal of Social Psychology, 40*(7), 1302–1309.
Sriram, N., & Greenwald, A. G. (2009). The brief Implicit Association Test. *Experimental Psychology, 56*(4), 283–294.
Sumner, W. G. (1906). *Folkways: A study of the sociological importance of usages, manners, customs, mores, and morals.* Boston: Ginn.
Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations, 33*, 33–47.
Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated-measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement, 72*(1), 37–43.
Trafimow, D. (2015). A defense against the alleged unreliability of difference scores. *Cogent Mathematics, 2*(1), 1064626.
Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria redefining merit to justify discrimination. *Psychological Science, 16*(6), 474–480.
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.
Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*(1), 59–69.