Manuscript in press, *Journal of Personality and Social Psychology*

Reducing Discrimination: A Bias Versus Noise Perspective

Jordan R. Axt
Duke University

Calvin K. Lai
Washington University in St. Louis

Word Count: 17,039

Author Contact Information
Jordan Axt
Duke University
Social Science Research Institute
334 Blackwell St #320
Durham, NC 27701
jordan.axt@duke.edu

Abstract

Discrimination can occur when people fail to focus on outcome-relevant information and incorporate irrelevant demographic information into decision-making. The magnitude of discrimination then depends on 1) how many errors are made in judgment and 2) the degree to which errors disproportionately favor one group over another. As a result, discrimination can be reduced through two routes: reducing noise -- lessening the total number of errors but not changing the proportion of remaining errors that favor one group -- or reducing bias -- lessening the proportion of errors that favor one group but not changing the total number of errors made. Eight studies (N = 7,921) investigate how noise and bias rely on distinct psychological mechanisms and are influenced by different interventions. Interventions that removed demographic information not only eliminated bias, but also reduced noise (Studies 1a-1b). Interventions that either decreased (Studies 2a-2c) or increased (Study 3) the time available to evaluators impacted noise but not bias, as did interventions altering motivation to process outcome-relevant information (Study 4). Conversely, an intervention asking participants to avoid favoring a certain group impacted bias but not noise (Study 5). Finally, a novel intervention that both asked participants to avoid favoring a certain group and required them to take more time when making judgments impacted bias and noise simultaneously (Study 5). Efforts to reduce discrimination will be well-served by understanding how interventions impact bias, noise, or both.

Word count: 232

Keywords: Bias, accuracy, noise, discrimination, signal detection, race, physical attractiveness

Reducing Discrimination: A Bias Versus Noise Perspective

Conceptualizing, measuring, and reducing discrimination has been a continual focus of social psychological research, and for good reason. Discrimination on the basis of characteristics like race, ethnicity, or gender has been at the center of political issues ranging from disproportionate police use of force against Black Americans in the United States (Cobb, 2016), hiring discrimination against women in the sciences (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012), and housing discrimination against Muslims and Roma across Europe (EU-MIDIS, 2017). Here, we consider discrimination as behavioral differences in treatment based on group membership. While stereotypes (mental associations between groups and characteristics) and prejudices (affective associations about groups; Fiske, 1998) are some of the mental *inputs* that contribute to discrimination, discrimination is the actual *output* for creating and maintaining real-world disparities in group outcomes.

Discrimination can take many forms. Often, discrimination is structural: existing rules, norms, or institutions preclude certain groups from access to the same rights or opportunities as other groups (Green, 2003). Here, discrimination can arise even if individual decision-makers do not incorporate demographic information into judgment (e.g., racial discrimination in college admissions as a result of minority applicants being disproportionately likely to attend lower-funded high schools and therefore achieving weaker standardized test scores). But in many other cases, discrimination occurs when demographic information is actively used in evaluation. This form of discrimination is evident when certain groups are disproportionately likely to receive positive outcomes than other groups, all else equal. Field-based audit studies, which seek to manipulate only targets' demographic information, have consistently shown such discrimination based on information like race or gender. For instance, when trained actors bargained for the

same model of car at the same dealership and followed the same script, White bargainers were offered greater discounts than Black bargainers (Ayres & Siegelman, 1995). Studies using similar methods have found that consequential judgments like those concerning admissions (Milkman, Akinola & Chugh, 2012), hiring (Neumark, 1996), housing (Ross & Turner, 2003) or lending (Ross & Yinger, 2002) are impacted by targets' demographic information.

Decisions in these domains are usually not simple, and frequently require considering multiple pieces of information to determine an appropriate response. For example, an admissions officer must decide if an applicant deserves to be accepted by integrating relevant criteria (e.g., GPA, extracurriculars, standardized test scores), while ignoring ostensibly irrelevant demographic information. These two aspects of the decision-making context -- the use of multiple pieces of relevant information and the presence of demographic information-- complicate judgment. For one, synthesizing across multiple relevant criteria is not straightforward. The admissions officer must weigh information that varies considerably on dimensions like scale (test scores vs. GPA) and format (e.g., recommendation letter versus personal essay). One consequence of this need to integrate across relevant criteria is that clear decision rules (e.g., only accepting applicants with very strong recommendation letters) are impossible or exceptionally difficult to create.

Indeed, prior research finds that synthesizing across multiple criteria creates ambiguity in determining an appropriate response, leading evaluators to use more heuristic thinking to simplify the decision-making process (Dhami, 2003). Rather than trying to maximize performance, decision-makers who must synthesize multiple criteria rely on "fast and frugal" strategies that balance the tradeoff between accuracy and efficiency (Dhami & Ayton, 2001; Dhami & Harries, 2001; Luan, Schooler & Gigerenzer, 2011). That is, individuals consider

outcome-relevant criteria until they arrive at an internal threshold where they feel comfortable making a decision. This threshold allows for "good enough" performance, where decision-makers are willing to accept the presence of errors in evaluation if they determine that the added benefit of increased accuracy is not worth the extra effort needed to achieve it. For instance, admissions officers will review each application long enough to get a general impression of whether the applicant deserves to be admitted or rejected, but not so long as to be unable to move through the large number of applications efficiently. The end result is that the admission officer will settle on a decision-making strategy that allows for satisfactory but far from perfect performance (Martignon, Katsikopoulos & Woike, 2008).

The presence of demographic information also alters the decision-making process. Most obviously, demographic information creates the opportunity for discrimination. The likelihood of discrimination is only further increased by the ambiguity created from having to integrate across multiple outcome-relevant criteria, as evaluators may incorporate demographic information into judgment when the correct decision is unclear (Dovidio & Gaertner, 2000). In addition, the presence of demographic information can directly reduce attention to outcome-relevant criteria, specifically when demographic information is not diagnostic (e.g., Feingold, 1992). Prior research has found that evaluators often falsely incorporate non-diagnostic information into judgment, and do so at the expense of using more diagnostic information (e.g., Castellan, 1973; Troutman & Shanteau, 1977). For instance, admissions officers may use demographic information like race or gender when forming an evaluation of an applicant, and doing so may lessen the attention paid to relevant criteria like grades. As a result, the presence of irrelevant demographic information not only allows for discrimination to be possible, but it can detract from attention that would have been otherwise given to parsing outcome-relevant information.

Using Signal Detection Theory to Understand Discrimination

In many contexts, decision-makers are susceptible to errors in judgment due to a) difficulty in correctly integrating across multiple criteria, and b) use of non-diagnostic demographic information. In this work, we apply signal detection theory (SDT; Green & Swets, 1966) to illustrate how these two influences can be used to understand the impact of discrimination, following previous research using SDT to explore other phenomena in social psychology like stereotyping (Park & Banaji, 2000) or dyadic relationships (Xu & Shrout, 2018)

Much of SDT's application in psychological research centers on how people make decisions under uncertainty. In these cases, evaluators must work to detect underlying populations in the presence of distracting information. For example, the college admissions officer strives to identify the worthy versus the unworthy applicants, and must do so using information that is difficult to parse accurately (e.g., deciding whether a glowing reference letter compensates for low grades).

An SDT approach to decision-making considers the degree of *noise* and *bias* in judgment. *Noise* refers to how well individuals perform at differentiating between the underlying populations in evaluation; less noise indicates that an individual has done a better job of detecting the signal (e.g., a superior job of admitting the more qualified applicants and rejecting the less qualified applicants). For analyses, we investigate the level of noise in judgment by calculating *sensitivity*, which is the degree to which an evaluator lacks noise; more sensitivity indicates less noise.

Bias, at least when applied to discrimination, is the extent to which a certain response is more likely for one group relative to another. For instance, a racially biased admissions officer may be more likely to accept White versus Black applicants. For analyses, we investigate the

presence of bias by calculating *criterion*-- the degree to which a particular response is more likely-- separately for targets from different demographic groups. Bias is then evident when criterion differs between targets from different demographic groups. Greater differences in criterion mean more bias. In much of the psychological literature, *bias* has been used synonymously with discrimination to indicate socially-based favoritism in behavior or treatment. Here, bias refers to the more specific outcome of group-level differences in response criterion.

Sensitivity and criterion are conceptually distinct; knowing one tells us very little about the other. For instance, evaluators can have the same level of criterion but differ in sensitivity. One admissions officer may be better than another at identifying the best-qualified applicants (greater sensitivity), but the two admissions officers could have the same levels of criterion if they falsely admit the same *proportion* of less-qualified applicants (i.e., when errors are made, both officers are equally more likely to commit errors of falsely accepting less qualified applicants than falsely rejecting more qualified applicants). Conversely, individuals can have the same level of sensitivity but differ in criterion. For instance, two admissions officers may be equally capable at identifying the more qualified applicants, but one admissions officer may be more likely to make errors that falsely admit less-qualified applicants (low criterion for giving an "accept" response) while another more likely to make errors that falsely reject more-qualified applicants (high criterion).

In these conditions, the amount of discrimination can be conceptualized by considering the number and type of errors made when evaluating members of different demographic groups. Specifically, the magnitude of discrimination can be determined by the degree to which the evaluation process simultaneously produces noise (i.e., high versus low sensitivity) and bias (i.e., large versus small differences in criterion between targets belonging to different demographic

groups). As a result, the magnitude of discrimination can change even if one of these factors remains constant; for example, discrimination can increase when more errors are made in evaluation (i.e., there is greater noise) even if bias (the rate at which those errors favor one group over another) does not change. Below, we provide a more specific example concerning how these two factors provide distinct pathways for reducing discrimination.

Reducing Discrimination via Bias or Noise: A Test Case

Consider a tech company concerned about gender discrimination in hiring. In a simplified example, suppose the company recently reviewed 100 applicants (50 male, 50 female) for 50 intern positions. Each applicant was presented with outcome-relevant information (e.g., coding experience, educational background) and gender information through a name on the resume. It is the organization's job to integrate across the relevant information to arrive at an overall evaluation of each applicant. In this simplified example, we assume that 50 applicants are more qualified (based on whatever weighting of outcome-relevant information) and should be hired, whereas 50 applicants are less qualified and should be rejected. Finally, suppose that of the 50 more qualified applicants, half are male and half female, and the same is true of the 50 less qualified applicants. In other words, gender is non-diagnostic; knowing an applicant's gender tells you nothing about whether they should be hired or rejected.

Given that hiring decisions should be based on outcome-relevant information, the treatment of each applicant can be coded as either "correct" (hiring a more qualified applicant or rejecting a less qualified applicant) or "incorrect" (hiring a less qualified applicant or rejecting a more qualified applicant). Note that, from the applicant's perspective, one type of error results in a beneficial outcome (being hired when less qualified) and another type of error results in a detrimental outcome (not being hired when more qualified).

In this case, gender discrimination can only occur when two factors are present: 1) the existence of errors (or noise) in the selection process and 2) an unequal distribution of errors (or bias), such that one group is more likely to receive "beneficial errors" and another group relatively more likely to receive "detrimental errors." For example, imagine if the company has 60% accuracy for the 50 applicants from each gender (30 correct decisions, 20 errors). However, the types of errors are not distributed evenly. For men, 75% of the errors are beneficial (15 less qualified men incorrectly hired) and 25% are detrimental (5 more qualified men incorrectly rejected). For women, the reverse is true; 25% of errors are beneficial (5 less qualified women incorrectly hired) and 75% of errors are detrimental (15 more qualified women incorrectly rejected). In this case, the distribution of hires and rejections would be:

	Males		Females	
	More Qualified	Less Qualified	More Qualified	Less Qualified
Hires	15	15	15	5
Rejections	5	15	15	15

Though men and women did not differ in overall qualifications, the combination of noise and bias resulted in discrimination, such that 60% of the hires were men.

How could the company reduce discrimination? One way is to make evaluators more accurate (i.e., reduce noise). Suppose an intervention helped the organization become considerably better at parsing an applicant's outcome-relevant information, such as by identifying what criteria were actually most related to job success and more actively stressing those criteria when selecting interns. Imagine that this intervention increased accuracy for male and female applicants from 60% to 92%. However, the intervention did not impact bias; 75% of

remaining errors towards men were still beneficial and 75% of errors towards women were still detrimental. The distribution of hires and rejections would then be:

	Males		Females	
	More Qualified	Less Qualified	More Qualified	Less Qualified
Hires	23	3	23	1
Rejections	1	23	3	23

Raising accuracy from 60% to 92% reduced noise and as a result reduced discrimination. Male applicants were still slightly overrepresented in the percentage of hires (52%) versus female applicants (48%), but the size of the gender gap was cut by 80%.

Another way to reduce discrimination is to lessen bias. Suppose a different intervention had no impact on accuracy (60%), but helped participants be more vigilant about gender bias, such as by warning evaluators to be on guard about possible gender-based favoritism in their judgments. The result of the intervention was that now only 55% of the errors towards men were beneficial and 45% of the errors towards women were beneficial. The distribution of hires and rejections would then be:

	Males		Females	
	More Qualified	Less Qualified	More Qualified	Less Qualified
Hires	15	11	15	9
Rejections	9	15	11	15

Altering the distribution of errors lessened bias and also reduced discrimination, such that again male applicants constituted only slightly more hires (52%) than female applicants (48%). This approach to reducing discrimination accomplished the same outcome as reducing noise but

through different means. Whereas the noise reduction approach worked by reducing the number of errors made, the bias reduction approach worked by changing the type of errors made.

Relations Between Bias and Noise

It is tempting to think that noise and bias in social judgment rely on similar processes. The mechanisms behind how many errors are made in evaluation are potentially quite related to how those errors are distributed between groups. Indeed, attending to relevant information is almost synonymous with ignoring irrelevant information in everyday language. And yet, bias and noise are empirically independent (Green & Swets, 1966). Knowing the amount of noise in judgment tells us almost nothing about the amount of bias. Given the analytical independence between noise and bias, the present work explores the possibility that the levels of noise and bias in judgment are determined by two distinct psychological processes; specifically, that the degree of noise in judgment is related to use of outcome-relevant information, whereas the degree of bias is related to use of irrelevant demographic information.

Knowing whether interventions effectively reduce discrimination via changes in bias versus noise has great practical and theoretical value. When judgments have a large amount of noise (i.e., difficult to distinguish between underlying populations) but bias is weak (i.e., relatively small differences in errors favoring one group over another), it may be easier to reduce discrimination by reducing noise rather than bias. For instance, evaluations of research grant proposals are notoriously noisy, with one analysis finding that the median single-rater reliability of grant quality was .33 (Cicchetti, 1991). Though the magnitude of bias is unknown, evaluations of research grants may exhibit racial discrimination, as a correlational analysis found that African American investigators were 13% less likely to receive NIH funding compared to Whites (Ginther et al., 2011). In noisy judgment contexts like grant evaluations, racial

discrimination may then be more effectively reduced from interventions that seek to reduce noise (for instance, providing more detailed instructions on scoring of proposal weaknesses; Sattler, McNight, Naney & Mathis, 2015) than interventions that seek to reduce bias, such as by asking reviewers to directly guard against any possible race-based favoritism in their evaluations.

Unfortunately, prior research has often relied on outcome measures unable to distinguish between the relative impact of bias and noise in discrimination (e.g., Bertrand & Mullainathan, 2004; Bodenhausen, Kramer & Susser, 1994; Uhlmann & Cohen, 2005). For example, Bohnet, van Geen and Bazerman (2015) examined the impact of a single evaluation (i.e., viewing partners one at a time) versus joint evaluation (i.e., viewing pairs of partners, one male and one female) judgment context on gender discrimination in choosing partners for a hypothetical academic competition. Participants had to select one partner from a pool of options, and partners were presented with relevant information (past academic performance) and ostensibly irrelevant demographic information (gender).

When participants completed the task in conditions of single evaluation, they were more likely to select male partners, even though more qualified female partners were available. When participants completed the task in conditions of joint evaluation, participants showed no gender-based discrimination. However, in this case, it's unclear whether joint evaluation reduced gender-based discrimination through greater attention to the relevant, academic information (reduced noise), or through greater ability to disregard the irrelevant gender information (reduced bias), or through both processes simultaneously. Understanding the relative contribution of bias and noise in discrimination requires measures that can distinguish them.

The Present Work

Using a novel task, we investigate whether different interventions impact the magnitude of discrimination in social judgment either by altering the degree of noise or bias. This work differs from past investigations using similar outcome measures (e.g., Axt, Casola & Nosek, 2018; Axt, Ebersole & Nosek, 2016; Axt & Nosek, 2018) by distinguishing between noise and bias as distinct routes for reducing discrimination. For example, Axt and Nosek (2018) investigated the necessary components required for interventions to effectively reduce criterion bias. They found that interventions needed to specifically mention the social category on which targets differed (i.e., physical attractiveness) compared to using more general information (i.e., warning participants that targets will differ on "irrelevant characteristics"). While such work advances our understanding of *what* is needed for certain interventions to effectively reduce criterion bias, the present work examines a broader issue of how advances in reducing the magnitude of discrimination can be achieved by lowering bias, noise, or both.

Here, we focus on two primary determinants of bias or noise in evaluation: 1) an ability to change the behavior most responsible for bias or noise, and 2) a motivation to do so. These two forces-- ability and motivation-- have been consistently highlighted as necessary in models of behavior change (Wilson & Brekke, 1994; Wegener & Petty, 1996), associations between attitudes and behavior (Fazio, 1990), and prejudice regulation (Monteith, 1993; Burns, Monteith & Parker, 2017). So long as people hold some level of ability and motivation to reduce noise or bias in judgment, then interventions targeting one of these factors should be effective at changing the relevant behavior (e.g., assuming people have any motivation to decrease the degree of noise in their evaluations, then increasing their ability to do so will result in less noise).

Across eight studies using a range of outcomes, social groups, and manipulations, we apply signal detection analyses to differentiate between bias and noise in discriminatory behavior. In Studies 1a and 1b, participants completed a judgment task either with or without the presence of irrelevant demographic information, and found that the presence of demographic information increased bias and noise. In the remaining studies, we tested whether various interventions differentially impact the degree of noise versus bias in judgment. Interventions that targeted one's ability or motivation to process outcome-relevant information -- such as by imposing time pressure (Studies 2a-2c), requiring delays in response (Study 3), or instilling greater motivation to engage in heuristic versus systematic thinking (Study 4)-- impacted noise but not bias. Conversely, an intervention that alerted participants to the social dimension responsible for favoritism in judgment impacted bias but not noise (Study 5). Finally, an intervention impacted bias and noise simultaneously by both alerting participants to the social dimension responsible for favoritism and requiring a delay in responding (Study 5). Together, these results suggest that bias and noise are distinct components of discriminatory behavior, and that the magnitude of discrimination can be reduced by targeting either or both outcomes.

Studies 1a and 1b

Studies 1a-1b tested several of the basic assumptions concerning how individuals form judgments when integrating across multiple pieces of information in the presence of non-diagnostic demographic information. Specifically, Studies 1a and 1b examine whether 1) people make decision-making errors even when no irrelevant information is provided, 2) when irrelevant demographic information is available, errors are increased (either by reduced attention to outcome-relevant criteria or increased attention to demographic information), and 3)

demographic information introduces bias by making certain types of errors more likely for some social groups than others.

Participants in these studies completed a version of the Judgment Bias Task (JBT; Axt, Nguyen & Nosek, 2018), which is designed to measure discrimination in social judgment. In this version of the JBT, participants evaluated applicants for an honor society based on relevant academic qualifications (GPA, recommendation letters, etc.). Participants either completed a 'blinded' JBT, where applicants were only shown with academic qualifications, or a JBT where applicants were also shown with social information known to influence judgment: physical attractiveness (Study 1a) or political affiliation (Study 1b). JBT performance was analyzed both in terms of overall noise (i.e., sensitivity) and bias (i.e., differences in criterion based on social group membership).

Method

Participants

For all studies, participants came from the Project Implicit research pool (implicit.harvard.edu; Nosek, 2005). In Study 1a, 911 participants (61.7% female, 71.1% White, *M*Age = 34.1, *SD* = 14.7) completed at least the JBT. In Study 1b, 636 participants (65.3% female, 77.8% White, *M*Age = 33.6, *SD* = 15.7) completed at least the JBT and reported being either a Democrat or Republican. Participants were only eligible for Study 1b if they reported being US citizens and either politically conservative or liberal (i.e., not neutral) when registering for the pool.

Study 1a sample size provided at least 80% power for finding a small between-subjects effect of Cohen's $d = .20$, and Study 1b sample size provided at least 80% power at detecting the effect size of blinding on sensitivity found in Study 1a ($d = .26$). For all studies, sample sizes

vary across tests due to missing data. See https://osf.io/mxn5b/ for Study 1a's pre-registration and https://osf.io/59a3p/ for Study 1b's pre-registration. Materials, data, and analysis scripts for all studies can be found at https://osf.io/brg76/. The online supplement, which includes additional pre-registered analyses, can be found at https://osf.io/ewqka/.

Procedure

Participants completed the JBT, followed by measures of perceived performance, desired performance, explicit preferences and implicit associations.

Academic Judgment Bias Task. In both studies, participants completed an academic Judgment Bias Task (JBT; Axt, Nguyen & Nosek, 2018). Participants received instructions that they would be making accept or reject decisions for applicants to a hypothetical academic honor society. Each application contained four pieces of information (Science GPA, Humanities GPA, Letter of recommendation quality, and interview score; see Appendix A), and participants were instructed to weight each piece of information equally. Qualifications were manipulated to create two levels of applicant quality so that there were "correct" and "incorrect" decisions. Half of the applications were scored to be equally less qualified, and half were scored to be equally more qualified (see Axt et al., 2018, Study 1a for scoring details). The JBT consisted of 64 unique profiles, and participants were instructed to accept approximately half of the applicants. Before making their accept and reject decisions, participants first viewed each application one at a time for one second each during an encoding phase.

In both studies, participants in the *Blind* condition saw only the applications with no additional information. In Study 1a, participants in the *Unblind* condition viewed the same applications with a face that was pre-rated as either more or less physically attractive (d = 2.64; Axt, Nguyen & Nosek, 2018). There were equal numbers of males and females within each level

of physical attractiveness. In Study 1b, participants in the *Unblind* condition viewed applications paired with an image indicating political affiliation (a donkey logo for Democrat and an elephant logo for Republicans). In both *Unblind* conditions, the more and less qualified applications were split evenly among more and less physically attractive people or Democrats and Republicans.

In Study 1a, participants in the *Unblind* condition were randomly assigned to one of 16 orders. Across orders, each face was equally likely to be assigned to a more or less qualified application. In Study 1b, participants in the *Unblind* condition were randomly assigned to one of 12 orders, with each application being equally likely to be described as a Democrat or Republican across orders.

Performance, explicit attitude and identity measures. Following the JBT, participants in *Unblind* conditions completed two items assessing perceived and desired task performance. Participants first reported their perceived performance for treating applicants from the study's two social groups (e.g., -3= "I was extremely easier on less physically attractive applicants and tougher on more physically attractive applicants", +3' "I was extremely easier on more physically attractive applicants and tougher on less physically attractive applicants"), followed by their desired performance (e.g., -3= "I wanted to be extremely easier on Republican applicants and tougher on Democrat applicants", +3= "I wanted to be extremely easier on Democrat applicants and tougher on Republican applicants").

Participants in both studies also reported their explicit preferences for the study's two social groups (e.g., in Study 1a, -3= "I strongly prefer less physically attractive people to more physically attractive people", +3= "I strongly prefer more physically attractive people to less physically attractive people").

Finally, participants in Study 1b reported their political identification (Democrat, Republican, Independent, Libertarian, Green, Other, Do not know). If participants selected an option other than Democrat or Republican, they then answered a forced-choice item asking them which of the two parties they would identify with if they had to. To maximize power, we grouped participants as Democrats or Republicans if they selected that party on either item, given prior work that such participants behave similarly in political judgment (Hawkins & Nosek, 2012). For Study 1b, data were also analyzed by whether applicants were from the same or opposing political party to maximize power.

Implicit associations. Participants in Study 1a completed a seven-block Implicit Association Test (IAT; Greenwald, McGhee & Schwartz, 1998) assessing implicit evaluations of more versus less physically attractive people, with stimuli coming from separate faces pre-rated to vary in physical attractiveness (Ma, Correll & Wittenbrink, 2015). Participants in Study 1b completed an IAT measuring implicit identification with Democrats and Republicans (e.g., liberal, conservative, Barack Obama, George Bush). In all studies, IATs were scored by the *D* algorithm (Greenwald, Nosek & Banaji, 2003). Higher values indicated more positive implicit evaluations of more physically attractive people and greater identification with Democrats.

Results

Participants were excluded from analysis for accepting fewer than 20% or more than 80% of applicants, with participants in *Unblind* conditions also being excluded for accepting (or rejecting) all applicants from either social group. Among those completing the JBT, these criteria resulted in 5.0% of Study 1a participants and 4.4% of Study 1b participants being excluded.[1]

[1] JBT exclusion rates did not reliably differ between conditions in Studies 1a-1b and 3-5, but did differ in Studies 2a-2c. The online supplement provides additional analyses as robustness checks for Studies 2a-2c as well as exclusion rates for all conditions and all studies.

Across studies, participants were also excluded for analyses involving the IAT if more than 10% of responses on critical trials were faster than 300 milliseconds (Nosek, Greenwald & Banaji, 2005; 1.7% of additional participants in Study 1a and 1.5% in Study 1b).

Sensitivity and Bias in Decision-Making

Accuracy (rate of accepting more qualified applicants and rejecting less qualified applicants) in all conditions was above chance (all t's > 42.47, all p's <.001). We calculated sensitivity and criterion using the same guidelines as Correll et al., 2007. See Table 1 for descriptive statistics for overall accuracy, sensitivity, and criterion for each social group across all conditions in Studies 1a-1b.

In Study 1a, sensitivity was lower in the *Unblind* than in the *Blind* condition, $t(863) = 3.87$, $p < .001$, $d = .26$ [.13. .40], and the same was true in Study 1b, $t(606) = 2.26$, $p = .024$, $d = .18$ [.02. .34]. In both studies, participants in *Unblinded* conditions showed bias, meaning differences in criterion between social groups (replicating Axt et al., 2018). In Study 1a, criterion was lower for more versus less physically attractive applicants, $t(430) = 5.19$, $p < .001$, $d = .25$ [.15. .35]. In Study 1b, criterion was lower for applicants from one's own versus the other political party, $t(319) = 4.62$, $p < .001$, $d = .26$ [.15. .37].

Associations with Attitude and Performance Measures

See Table 2 for descriptive statistics for perceived performance, desired performance, explicit preferences and implicit associations across all conditions and studies.

For all studies, we tested for differences across conditions in perceived performance, desired performance, explicit preferences, and implicit associations. These analyses failed to produce consistent effects (e.g., only 3 out of the 28 analyses produced reliable differences at $p <$

Table 1

Means (and standard deviations) for overall JBT accuracy, sensitivity, and criterion for each social group in Studies 1a-1b

Study 1a Condition	JBT Accuracy	Sensitivity	More Attractive Criterion	Less Attractive Criterion
Blind (*N* = 434)	69.65% (7.83)	1.15 (.52)		
Unblind (*N* = 431)	67.52% (8.22)	1.01 (.52)	-.13 (.46)	-.01 (.45)
Study 1b Condition	JBT Accuracy	Sensitivity	Ingroup Criterion	Outgroup Criterion
Blind (*N* = 288)	70.65% (8.04)	1.22 (.53)		
Unblind (*N* = 320)	69.34% (8.15)	1.12 (.53)	-.15 (.47)	-.01 (.44)

Table 2

Means (and standard deviations) for perceived performance, desired performance, explicit preferences and implicit associations

Study 1a Condition	Perc. Performance	Des. Performance	Exp. Preferences	Imp. Associations
Blind			1.87 (1.14)	.44 (.42)
Unblind	.19 (.71)	.16 (.75)	1.80 (1.06)	.36 (.46)
Study 1b Condition				
Blinde			1.06 (1.03)	.75 (.36)
Unblind	.07 (.77)	-.01 (.57)	.74 (.99)	.74 (.39)
Study 2a Condition				
High Time Pressure	.09 (.85)	-.13 (.59)	.28 (.63)	.26 (.41)
Moderate Time Pressure	.09 (.71)	-.06 (.65)	.18 (.68)	.29 (.48)
Low Time Pressure	.25 (.72)	-.02 (.40)	.31 (.72)	.26 (.48)
Study 2b Condition				
Timed	.13 (.84)	-.01 (.62)	.78 (1.02)	.65 (.40)
Untimed	.10 (.70)	.02 (.52)	.79 (.89)	.69 (.41)
Study 2c Condition				
Timed	.15 (.89)	.01 (.62)	.79 (.97)	.66 (.40)
Untimed	.10 (.79)	.02 (.58)	.84 (.95)	.70 (.42)
Study 3 Condition				
Control	.14 (.71)	-.02 (.39)	.83 (.94)	.68 (.37)
Delay	.12 (.75)	0 (.43)	.72 (.97)	.70 (.37)
Study 4 Condition				
Heuristic	.12 (.78)	-.03 (.55)	.86 (.96)	.71 (.39)
Control	.08 (.71)	-.002 (.46)	.80 (.90)	.72 (.37)
Systematic	.08 (.58)	-.01 (.43)	.78 (.91)	.71 (.38)
Study 5 Condition				
Control	.10 (.73)	0 (.56)	.79 (.99)	.70 (.39)
Bias Warning	-.01 (.56)	-.04 (.47)	.72 (.93)	.68 (.38)
Bias Warning + Delay	-.04 (.63)	-.03 (.55)	.68 (1.01)	.69 (.39)

Note. Perc. Performance = Perceived JBT performance. Des. Performance = Desired JBT performance. Exp. Preferences = Explicit preferences (Attractiveness for Studies 1a and 2b-5), political ingroup (Study 1b) or race (Study 2a). Imp. Associations = Outcome on

measure of implicit associations (IAT assessing attractiveness attitudes in Studies 1a and 2b-5, IAT assessing political identification in Study 2b, BIAT assessing racial attitudes in Study 2a).

.05, and these effects failed to replicate across studies). Specific results for each study can be found in the online supplement.

We also analyzed whether criterion biases were associated with perceived performance, desired performance, explicit preferences and implicit associations. We present these analyses in the aggregate across all studies here, but results for individual studies are available in the online supplement. Replicating prior work (Axt, Nguyen & Nosek, 2018; Axt, Casola & Nosek, 2018; Axt, Ebersole & Nosek, 2016), criterion biases were modestly but reliably associated with desired performance (r = .10, 95% CI[.02, .19]), perceived performance (r = .22, 95% CI[.14, .29]), explicit preferences (r = .11, 95% CI[.09, .14]), and implicit associations (r = .09, 95% CI[.06, .11]).

Discussion

Participants who completed an evaluation task in the presence of non-diagnostic demographic information made more errors (i.e., more noise and lower sensitivity). These errors were also biased (i.e., differences between social groups in response criterion), such that some groups were more likely to receive beneficial treatment than others (more physically attractive people in Study 1a and political ingroup members in Study 1b). On average, across the 64 total judgments, removing demographic information in Study 1a led to 2.7 fewer errors (i.e., the number of qualified applicants incorrectly rejected or less qualified applicants incorrectly accepted) and 1.7 fewer errors in Study 1b.

These results shed light on how individuals use outcome-relevant information and irrelevant demographic information in judgment. First, though noise was lower in *Blind* conditions, that did not mean performance was without error (Study 1a *Blind* error rate = 30.5%, Study 1b *Blind* error rate = 29.4%). The JBT is designed for applicants to systematically differ in

qualification level based on only four pieces of relevant information, and it is conceivable for participants to "crack the code" behind the JBT's scoring and achieve 100% accuracy. However, this did not happen even when demographic information was removed. Rather, participants appeared to parse the outcome-relevant information so that they could achieve above-chance accuracy without spending too much time to do so. This suggests that some level of noise may be due to an inability or unwillingness to further parse outcome-relevant information.

That said, the presence of non-diagnostic demographic information increased noise. If participants did not incorporate the demographic information into judgment, there would have been no differences in sensitivity between *Blind* and *Unblind* conditions. Instead, the lower sensitivity in *Unblind* conditions suggests that the presence of demographic information increased noise by reducing attention to academic information in judgment, enabling the use of demographic information, or both.

Finally, bias was evident when demographic information was available. Participants were more likely to give members of more favored groups beneficial treatment (i.e., acceptance to the academic honor society) than members of less favored groups. The presence of irrelevant demographic information increased errors, but not in a way that simply made all errors equally more likely. Rather, the presence of demographic information increased errors in favor of preferred social groups.

Altering Noise versus Bias in Social Judgment

Studies 1a-1b illustrate how the presence of demographic information can increase noise and create bias in social judgment. In the remaining studies, we examine whether interventions differentially influence noise versus bias. Specifically, we test how noise changes due to the motivation or ability to process outcome-relevant information (Studies 2a-4), and how bias

changes due to the motivation or ability to avoid processing irrelevant demographic information (Study 5).

Time is a straightforward determinant of the ability to process outcome-relevant information. Many studies have found that optimal performance on judgment tasks decreases as time pressure increases (for review, see Ariely & Zakay, 2001). Decreasing the amount of time available to evaluators has been shown to be especially detrimental for judgments that rely on integrating across multiple pieces of information (Rothstein, 1986), and such time pressure is particularly effective at limiting the processing of outcome-relevant information (e.g., Keinan, 1987; McDaniel, 1990). In Studies 2a-2c, we examine the impact of added time pressure on both bias and noise in social judgment. Conversely, imposing time delays in judgment by requiring people to consider their judgments longer than they normally would has been found to increase accuracy and performance (Moulton, Regehr, Mylopoulos & Macrae, 2007), and in Study 3 we test the effect of requiring delays in responses. We would expect the amount of noise in decision-making to increase when people are placed under time pressure and decrease when people are compelled to take longer than normal to decide.

In Study 4, we examine how increasing reliance on systematic versus heuristic thinking affects noise and bias in social judgment. Motivation to process outcome-relevant information can be manipulated through inducing systematic versus heuristic thinking (Chaiken, 1980). For judgments that demand some deliberation, such as those requiring integration across multiple pieces of information, a more intuitive, heuristic mindset should be associated with increased noise while a more deliberate, systematic mindset associated with decreased noise (Kruglanski & Gigerenzer, 2011). Indeed, past work has found that engaging in more critical thinking can lead to more accurate evaluation. For instance, participants placed in a mindset of more systematic

thinking (by recalling a situation in which careful reasoning was helpful) showed more empathic accuracy (Ma-Kellams & Lerner, 2016). Other research involving more complex judgments (e.g., predicting financial performance of corporate bonds) has similarly found that accuracy increases with more deliberate processing, such as through heightening feelings of accountability (Ashton, 1992).

Conversely, efforts to increase one's ability or motivation to limit favoritism based on demographic information should reduce bias. In many contexts, heightening motivation to avoid favoritism in social judgment may be difficult to implement, as many people already report a strong desire to be unbiased. For instance, in Axt et al. (2018), 86.6% of participants completing a JBT reported not wanting to favor applicants based on physical attractiveness, and 84.5% reported not wanting to favor applications based on political affiliation. Manipulations that increase motivation to be unbiased may then be more effective for groups where individuals normally believe discrimination is acceptable (e.g., drunk drivers; Crandall, Eshleman & O'Brien, 2002). However, in contexts where motivation to be unbiased is high, reductions in bias should still follow from increasing ability to regulate the influence of demographic information on judgment. In Study 5, we test one method for doing so by drawing attention to the social dimension responsible for biased judgment (Pronin & Krugler, 2006). Alerting people to the presence of socially-based favoritism may allow them to effectively monitor their behavior, increasing control over bias in judgment and creating greater alignment with existing motivations to be unbiased.

Notably, interventions that increase ability to regulate bias need not translate into reduced noise (i.e., fewer errors) if bias is reduced through overcorrection (Dunton & Fazio, 1997; Sommers & Kassim, 2001). For instance, in Axt and Nosek (2018; Study 1), participants

completed the same attractiveness JBT as in Study 1a. Some participants were told beforehand that they would likely favor more physically attractive applicants and they should avoid doing so. Relative to a control condition, this manipulation reduced bias (i.e., relative differences in criterion for more versus less physically attractive applicants; d = .33), but a re-analysis found the manipulation had no reliable impact on noise (i.e., overall sensitivity on the JBT; d = .01). Participants simply became more likely to reject all physically attractive applicants and accept all less physically attractive applicants, regardless of actual qualifications. The end result was that the manipulation reduced discrimination, perhaps by increasing the ability to counteract bias based on demographic information, but the level of noise in judgment was unchanged. In Study 5, we use a similar manipulation that warned participants of a potential for favoritism towards more attractive people in social judgment, and examined whether the manipulation impacted bias but not noise.

Studies 2a-2c

In Studies 2a-2c, we tested how noise and bias in social judgment are influenced by the amount of time available to complete judgment. Past research strongly suggests that time pressure should increase errors, thereby decreasing sensitivity and increasing noise (Ariely & Zakay, 2001). It is less clear whether time pressure would impact bias (i.e., relative differences in criterion between social groups).

To consider how time pressure would impact bias, imagine a participant completing an attractiveness JBT without time pressure. When evaluating more physically attractive people, the participant makes two "detrimental" errors (falsely rejecting a qualified applicant) and four "beneficial" errors (falsely accepting a less qualified applicant). Suppose also that the reverse happens when evaluating less physically attractive people-- four detrimental errors and two

beneficial errors. In this case, 66% (4/6) of the errors towards more attractive people are beneficial compared to 33% (2/6) of the errors towards less attractive people, suggesting a strong bias favoring more physically attractive people.

Now, imagine that this participant completed the same attractiveness JBT but did so with time pressure. How would the results change? One possibility is that time pressure causes the participant to make additional errors, but those errors are evenly distributed across all error types, which would create a reduction in attractiveness bias. If time pressure caused 12 additional random errors on top of the errors that one would normally make, that would mean 55% (10/18) of the errors would be beneficial for more attractive people and 45% (8/18) of the errors would be beneficial for less attractive people. The introduction of evenly distriuted errors would reduce the disparity in beneficial errors from 33% with no time pressure to 10% with time pressure.

A second possibility from research on automaticity and biased judgment suggests that time pressure causes people to rely more on social heuristics to guide decision-decision-making (e.g., Gilbert & Hixon, 1991; Fazio, 1990). In that scenario, time pressure would increase reliance on physical attractiveness in decision-making, causing the participant to make additional errors that are skewed in favor of physically attractive participants. In other words, time pressure would increase attractiveness bias.

A final possibility is that time pressure will have no impact on bias. This would occur if time pressure increased the total number of errors, but these errors accrued in *the same proportion* as when evaluation occurred without time pressure. As time pressure reduces the ability to deliberate, such results would suggest that the mental processes underlying bias are automatic and do not depend on much deliberation. Drawing from the prior example, time pressure may still triple the amount of errors, but these errors could be distributed in the same

proportion as in conditions without time pressure (e.g., 18 errors towards more attractive people, with six detrimental errors and twelve beneficial errors, such that 66% of errors are still beneficial). In this case, the same degree of bias in social judgment would simply operate across a greater number of errors.

To examine the impact of time pressure on noise and bias, we experimentally manipulated the amount of time that participants had to make judgments in Studies 2a-2c. We investigated this question across two judgment tasks. In Study 2a, participants completed a First-Person Shooter Task (FPST; Correll, Park, Judd & Wittenbrink., 2002), which required making 'Shoot' or 'Don't Shoot' decisions when viewing Black and White targets holding either guns or benign objects. Typical results find that participants exhibit a racial bias in judgment, with a lower criterion to give a "shoot" response for Black versus White targets (e.g., Correll et al., 2002; Correll, Park, Judd & Wittenbrink, 2007; Mendoza, Gollwitzer & Amodio, 2010). In Studies 2b-2c, we used the same physical attractiveness JBT as in Study 1a.

Study 2a

Method

Participants

The study was restricted to non-Black participants, given prior evidence that Black participants show a smaller bias on the FPST than non-Black participants (Correll et al., 2002, Study 4). In total, 706 participants completed at least the FPST, though only 451 participants (56.1% female, 73.8% White, *M*Age = 33.3, *SD* = 12.2) met our exclusion criteria (see Results). This sample of eligible participants provided an average of 98.7% power at detecting a medium effect of Cohen's *d* = .50.

Procedure

Participants first completed the FPST, followed by measures of perceived performance, desired performance, explicit racial preferences and implicit race-danger associations.

First-Person Shooter Task. Participants completed an 80-trial FPST (Correll et al., 2002). In the task, participants view scenes that eventually display Black or White men holding either guns or harmless objects (e.g., wallets). Participants must press one button to 'Shoot' targets holding guns and another to 'Not Shoot' targets holding harmless objects. The 80 trials contained 20 scenes, each with four versions (a Black vs. White target holding a gun vs. object). Trials were presented in a random order. Before starting the 80-trial test block, participants completed a practice block of 16 randomly selected trials. Practice block data were not analyzed.

Participants were randomly assigned to an FPST where trials timed out at either 630 milliseconds (*High Time Pressure* condition)[2], 710 milliseconds (*Moderate Time Pressur*e condition), or 790 milliseconds per trial (*Low Time Pressure* condition). Participants were not given trial feedback, but were shown a 'Respond Faster' message following any timeouts.

Other measures. Participants completed the same measures of perceived performance and desired performance as Study 1a, adapted to deal with failing to shoot White versus Black people. Participants also reported their relative explicit racial preferences and completed a four-block Brief Implicit Association Test (BIAT; Sriram & Greenwald, 2009) measuring associations between White and Black people with the concepts of danger versus safety (see Appendix B for procedure details and exclusion criteria).

[2] Timing was selected based on previous work using the FPST, which had a 630-millisecond response window (e.g., Correll et al., 2007).

Results

Participants were excluded from analysis for failing to respond to at least 40 of the 80 FPST trials and failing to achieve at least 55% accuracy in completed trials. This stringent exclusion criteria was used to ensure that analyses focused on participants paying adequate attention to the demanding task, but meant a substantial proportion of the sample was excluded, and that exclusion rates differed across conditions (High Time Pressure = 50%; Medium Time Pressure = 36.6%; Low Time Pressure = 22.6%). In-text analyses use the pre-registered exclusion criteria, but the online supplement contains additional analyses as robustness checks, which used either all participants or various forms of less stringent exclusion criteria. These supplemental analyses replicated the results described below, with one exception: analyses with all participants found much weaker effects of an overall racial bias in criterion (i.e., lower criterion to give a shoot response for Black versus White targets).

Sensitivity and Bias in Decision-Making

Sensitivity differed across conditions, $F(2,448) = 30.17$, $p < .001$, $\eta^2p = .12$, 95% C.I. [.07, .17]. Follow-up comparisons revealed that less time pressure was related to greater sensitivity. Participants in the High Time Pressure condition had lower sensitivity than those in the Moderate Time Pressure condition, $t(264) = 4.56$, $p < .001$, $d = .57$, 95% CI [.32, .81], and even lower sensitivity than those in the Low Time Pressure condition, $t(295) = 7.90$, $p < .001$, $d = .95$, 95% CI [.70, 1.19]. In turn, participants in the Moderate Time Pressure condition had lower sensitivity than those in the Low Time Pressure condition, $t(337) = 3.43$, $p = .001$, $d = .37$, 95% CI [.16, .59]. See Table 3 for descriptive statistics for overall accuracy, sensitivity, and criterion for each social group across all conditions in Studies 2a-2c, as well as for within-subjects tests of differences in criterion for Black versus White targets.

Table 3

Means (and standard deviations) for overall JBT or FPST accuracy, sensitivity, and criterion for each social group in Studies 2a-2c

Study 2a Condition	FPST Accuracy	Sensitivity	Black Criterion	White Criterion	Criterion Bias
High Time Pressure (N = 112)	70.87% (9.56)	1.21 (.64)	-.01 (.46)	.10 (.40)	$t(111) = 3.29$, $p = .001$
Moderate Time Pressure (N = 154)	76.45% (11.26)	1.64 (.83)	.02 (.42)	.11 (.37)	$t(153) = 3.11$, $p = .002$
Low Time Pressure (N = 185)	80.63% (10.66)	1.96 (.87)	.04 (.36)	.10 (.35)	$t(184) = 2.04$, $p = .043$
Study 2b Condition	JBT Accuracy	Sensitivity	More Attractive Criterion	Less Attractive Criterion	Criterion Comp
Timed (N = 601)	61.16% (8.21)	.62 (.47)	-.15 (.43)	.02 (.44)	$t(600) = 8.14$, $p < .001$
Untimed (N = 699)	67.32% (8.67)	1.00 (.54)	-.13 (.46)	.04 (.48)	$t(698) = 9.15$, $p < .001$
Study 2c Condition					
Timed (N = 448)	59.63% (8.46)	.53 (.48)	-.20 (.44)	-.02 (.47)	$t(447) = 7.24$, $p < .001$
Untimed (N = 517)	66.11% (8.87)	.92 (.54)	-.18 (.46)	.01 (.45)	$t(516) = 8.29$, $p < .001$

Note. Criterion Bias = Within-subjects test comparing criterion values.

Racial bias (a difference score between criterion for Black versus White targets) did not differ across conditions, $F(2,448) = 0.96$, $p = .384$, $\eta^2p = .004$, 95% C.I. [0, .02]. Follow-up comparisons affirmed that no conditions reliably differed in criterion bias (all t's < 1.30, all p's > .194, all d's < .16; see online supplement). All three conditions showed racial bias in decisions to shoot (Correll et al., 2002), with lower criterion for Black versus White targets (all t's > 2.04, all p's < .043, all d's > .15; see Table 3). Here, lower criterion values mean greater likelihood of making a 'Shoot' versus 'Don't Shoot' response, indicating a bias for Black targets to be shot relative to White targets regardless of whether targets were holding guns versus benign objects.

Discussion

Replicating past work using the FPST, participants displayed a racial bias in criterion, adopting a lower criterion to 'shoot' for Black versus White targets (Correll et al., 2002; Correll et al., 2007). Racial bias in criterion was similar regardless of how much time pressure there was. On average, when participants erred by indicating "shoot" towards someone holding a benign object, conditions showed similar rates at which these errors occurred for Black versus White targets (Low Time Pressure = 50.9%, Moderate Time Pressure = 52.7%, High Time Pressure = 52.4%). Likewise, when falsely deciding to "not shoot" someone holding a gun, conditions showed comparable rates at which these errors occurred for White versus Black targets (Low Time Pressure = 52.5%, Moderate Time Pressure = 53.0%, High Time Pressure = 53.1%).

In contrast, increased time pressure resulted in more noise (i.e., lower sensitivity). More time pressure meant more errors (Low Time Pressure = 15.5 errors, Moderate Time Pressure = 18.9 errors, High Time Pressure = 23.4 errors). As errors were biased toward erroneously shooting Black versus White targets, or erroneously not shooting White versus Black targets, an increase in time pressure meant an increase in race-based discrimination in decisions to shoot.

In Studies 2b-2c, we replicated this design using another task (the JBT), another social domain (physical attractiveness), and more high-powered tests.

Studies 2b and 2c

Participants

In Study 2b, we sought to collect 1310 eligible participants who completed at least the JBT. This sample provided 96% power for detecting a small between-subjects effect (d = .20), and 80% power for detecting a between-subjects effect of d = .155, which is half the size of the typical bias in criterion favoring more physically attractive people on JBTs without a time constraint (d = .31; Axt, Nguyen & Nosek, 2018). In total, 1479 participants (67.0% female, 68.9% White, *M*Age = 32.3, *SD* = 14.1) completed at least the JBT. See https://osf.io/ykjx4/ for Study 2b's pre-registration.

In Study 2c, we sought to collect at least 900 eligible participants who completed at least the JBT. This sample size would provide greater than 85% power at detecting a small between-subjects effect of d = .20 and greater than 99% power at detecting the effect of time constraints on JBT sensitivity in Study 2b (d = .75). In total, 1133 participants (61.6% female, 61.4% White, *M*Age = 33.1, *SD* = 14.5) completed at least the JBT. See https://osf.io/efpzg/ for Study 2c's pre-registration.

Procedure

In both studies, participants first completed the JBT, followed by measures of perceived performance, desired performance, prejudice motivations, explicit preferences and implicit associations.

Academic JBT. Participants were randomly assigned to complete a timed or untimed JBT. In both studies, participants in the *Untimed* conditions completed the same JBT (Axt et al.,

2018) as participants in the *Unblind* condition of Study 1a (now using only 12 orders). In Study 2b, participants in the *Timed* condition completed the same task but with an 1800-millisecond response window (a pilot study using this response window found that 92% of participants could respond to at least 80% of trials). In Study 2c, participants in the *Timed* condition completed the task with a 1500-millisecond response window. In *Timed* conditions, a message to "please respond faster" appeared after any trial timeout. Participants in *Timed* conditions were warned about the time limit beforehand and instructed to respond to as many trials as possible.

Other measures. Participants completed the same measures of perceived performance, desired performance, explicit preferences, and implicit attractiveness evaluations as in Study 1a. Participants also completed the five-item Internal Motivation to Respond Without Prejudice scale (IMS; Plant & Devine, 1998), adapted to be about treatment of less physically attractive people (e.g., "I am personally motivated by my beliefs to be non-prejudiced toward less physically attractive people"; 1= Strongly disagree, 9= Strongly agree).[3]

Results

Participants in Untimed conditions were excluded from analyses using the same criteria as Unblind conditions in Study 1a, while participants in Timed conditions were also excluded from analysis if they did not respond to at least 80% of trials. This latter requirement resulted in differences in overall JBT exclusion rates between Timed and Untimed conditions (Study 2b: Untimed = 8.0%, Timed = 15.5%; Study 2c: Untimed = 8.0%; Timed = 21.5%). In-text analyses use the pre-registered exclusion criteria, but the online supplement contains analyses using all participants. In both studies, conclusions do not change when including all participants.

[3] Greater internal motivation to avoid prejudice was weakly but reliably associated with lower biases in criterion favoring more physically attractive applicants (Study 2b r = -.09, 95% CI [-.15, -.04], Study 2c r = -.07, 95% CI [-.14, -.01]).

Sensitivity and Bias in Decision-Making

In both studies, participants in the Timed conditions had lower sensitivity than participants in the Untimed conditions, (Study 2b: $t(1298) = 13.53$, $p < .001$, $d = .75$, 95% CI [.64, .87]; Study 2c :$t(963) = 11.75$, $p < .001$, $d = .76$, 95% CI [.63, .89]). See Table 3 for descriptive statistics.

In both studies, conditions did not reliably differ in the size of criterion bias (a difference score between criterion for more and less physically attractive applicants). Participants in the Timed conditions showed no reliable differences from participants in Untimed conditions (Study 2b :$t(1298) = .01$, $p = .992$, $d = .001$, 95% CI [-.11, .11]; Study 2c :$t(963) = -.19$, $p = .847$, $d = -.01$, 95% CI [-.14, .11]). All conditions showed a physical attractiveness bias, with lower criterion for more versus less physically attractive applicants (t's > 7.24, p's < .001, d's > .33; see Table 3), meaning that more physically attractive applicants were more likely to be admitted to the academic honor society regardless of actual qualifications.

Discussion

Studies 2b-2c conceptually replicated Study 2a's results using a different outcome measure, social groups, and timing constraints. We found that time pressure on the JBT increased the magnitude of discrimination based on physical attractiveness through increased noise (lower sensitivity). Specifically, time pressure increased discrimination by raising the number of errors made in judgment (Untimed: Study 2b = 20.9 errors, Study 2c = 21.7 errors; Timed: Study 2b = 24.9 errors, Study 2c = 25.9 errors). However, time pressure had no impact on bias (relative differences in criterion between more and less physically attractive applicants), as all conditions showed comparable levels of lower criterion for more versus less physically attractive people. When less qualified applicants were incorrectly admitted, the rate at which

errors occurred for more versus less physically attractive people did not differ strongly across conditions (Untimed: Study 2b = 54.2%, Study 2c = 54.5%; Timed: Study 2b = 53.2%, Study 2c = 53.6%). When more qualified applicants were incorrectly rejected, the rate at which these errors occurred for more versus less physically attractive people also did not differ strongly across conditions either (Untimed: Study 2b = 54.0%, Study 2c = 54.1%; Timed: Study 2b = 54.1%, Study 2c = 54.5%).

Across Studies 2a-2c, time pressure did not strengthen the degree to which demographic information impacted the distribution of errors in judgment. Time pressure did not result in greater bias by increasing reliance on demographic information in judgment, nor did it create less bias by simply making all types of errors equally more likely. Rather, time pressure increased the probability of making errors in judgment, meaning time pressure caused the same propensity for bias to be expressed across a greater number of errors. These results suggest that manipulations like time pressure, which reduce individuals' ability to parse outcome-relevant information, are much more related to processes that determine the level of noise versus the level of bias in social judgment. These findings appear to be inconsistent with research on automaticity and biased judgment suggesting biases will increase with time pressure (e.g., Gilbert & Hixon, 1991; Fazio, 1990). We explore this apparent inconsistency in more detail in the General Discussion.

Study 3

Study 3 investigated how requiring participants to deliberate impacts the degree of bias and noise in evaluation. Participants completed an attractiveness JBT either at their own pace, or completed a writing task about why deliberating was important and then could only respond after

a 4500-millisecond delay.[4] While this manipulation is essentially the reverse of the time pressure manipulation used in Studies 2a-2c, it is possible that requiring a delay in response may produce different results. For one, mandatory response delays may have no impact on noise or bias if participants are unable or unwilling to use the added time effectively (i.e., through greater consideration of outcome-relevant criteria). Conversely, delayed responses could even translate into decreased noise and decreased bias if participants use the additional time to both better parse the outcome-relevant criteria *and* to counteract the influence of the biasing demographic information. Delayed responses could even increase bias if greater time led to the deliberate rationalization of biased reasoning (Kunda, 1990; Rand et al., 2014).

Finally, requiring delayed responses could have a similar (though opposing) effect as time pressure if added time reduces errors, but remaining errors continue to favor more over less physically attractive applicants at the same proportion as when judgments are made with no required delay. Such results would suggest that required delays in responding can be used effectively by participants to reduce errors in judgment, but the benefits of such delays are limited to processes determining noise rather than bias.

Method

Participants

We sought to collect at least 373 eligible participants who completed the JBT, which would provide greater than 80% power at detecting a small between-subjects effect of $d = .20$. The final sample was slightly larger. In total, 808 participants (67.6% female, 68.3% White,

[4] In the *Untimed* condition of Study 2c, 87% of participants had an average response time lower than 4500 milliseconds, meaning a large majority of participants would have already made their decisions by this point.

*M*Age = 32.4, *SD* = 13.7) completed at least the JBT. See https://osf.io/gq6sk/ for Study 3's pre-registration.[5]

Procedure

Participants first completed the JBT, then measures of perceived performance, desired performance, explicit preferences and implicit associations.

Academic JBT. Participants completed a modified version of the JBT used in Studies 2b-2c. Participants were randomly assigned to an *Untimed* condition that was the same as Studies 2b-2c or to a *Delay* condition, where participants were required to wait 4500 milliseconds before being able to make their 'accept' or 'reject' decision. Participants were warned ahead of time about the required delay in responding. In addition, participants in the Delay condition completed a short writing exercise before the JBT where they listed two reasons why thinking hard about one's decisions was important when completing the admissions task (see Appendix C). Exploratory analyses of how responses to the writing exercise related to JBT performance suggest that the manipulations' effects were driven by the time delay in responding, so our interpretation focuses on the impact of the time delay.[6]

[5] Study 3 used a *p*-critical design (Sagarin, Ambler & Lee, 2014). We checked our data at *N* =200 and *N* = 373 per experimental condition, noting that we would stop data collection at *N* = 200 per condition if bias in criterion *did* show reliable differences between conditions. We did not find this effect and therefore collected our full sample. As a result, our critical *p* value for rejecting the null hypothesis in Study 3 was $p < .0295$, which allows the study-wide Type I error rate to remain at 5%.

[6] We examined whether JBT performance was related to effort spent on the writing task, and found that it was not. The median character length for the writing exercise was 73, and the median time spent writing was 59 seconds. In exploratory analyses, we found that sensitivity was unrelated to number of characters produced in the writing exercise, either analyzed as raw (r = .04, p = .422), or log-transformed (r = .08, p = .124). We also found that sensitivity was not related to time spent on the writing task, either in raw (r = -.01, p = .823) or log-transformed (r = - .07, p = .178) seconds. Finally, we found that the experimental manipulation was effective at reducing sensitivity relative to Control participants even among those who wrote relatively few characters (i.e., only 50; $t(486) = 2.08$, p =.038) or spent relatively little time on the writing task (i.e., less than 35 seconds; $t(432) = 2.64$, p = .009).

Other measures. Participants completed the same measures of perceived performance, desired performance, explicit preferences and implicit associations as in Studies 2b-2c.

Results

Using the same criteria as Studies 2b-2c, 6.3% of participants were excluded from JBT analyses.

Sensitivity and Bias in Decision-Making

Participants in the Delay condition took longer to make each JBT decision (Average = 7095 ms; Median = 5639 ms) than participants in the Untimed condition (Average = 2900 ms; Median = 2677 ms).

See Table 4 for descriptive statistics for overall accuracy, sensitivity, and criterion for each social group across all conditions in Studies 3-5. Both conditions were above chance in accuracy (all t's > 38.75, all p's < .001). However, the Delay condition had greater sensitivity than the Untimed condition, $t(755) = 4.62$, $p < .001$, $d = .34$, 95% CI [.19, .48]. Both conditions also showed bias, with lower criterion for more over less physically attractive applicants (t's > 4.65, p's < .001, d's > .24; see Table 4), though the *Delay* and *Untimed* conditions did not reliably differ in levels of bias, $t(755) = 1.13$, $p = .260$, $d = .08$, 95% CI [-.06, .22].

Discussion

A mandatory delay in responding reduced discrimination on the basis of physical attractiveness. Specifically, requiring a response delay reduced discrimination by increasing sensitivity, and thereby making errors less likely (Control = 19.1 errors, Delay = 20.9 errors). However, response delays had no impact on bias. The experimental conditions did not show differences in the rate at which errors of falsely admitting less qualified applicants occurred towards more versus less physically attractive people (Control = 52.9%, Delay = 52.1%), or in

Table 4

Means (and standard deviations) for overall JBT accuracy, sensitivity, and criterion for each social group in Studies 3-5

Study 3 Condition	JBT Accuracy	Sensitivity	More Attractive Criterion	Less Attractive Criterion	Criterion Bias
Control (N = 382)	67.29% (8.72)	1.00 (.56)	-.11 (.45)	.02 (.46)	$t(381) = 5.51$, $p < .001$
Delay (N = 375)	70.01% (8.02)	1.18 (.53)	-.09 (.46)	.003 (.45)	$t(374) = 4.65$, $p < .001$
Study 4 Condition					
Heuristic (N = 490)	66.48% (8.05)	.93 (.49)	-.13 (.43)	.02 (.46)	$t(489) = 6.67$, $p < .001$
Control (N = 512)	67.51% (8.25)	1.01 (.52)	-.09 (.45)	.02 (.47)	$t(511) = 5.69$, $p < .001$
Systematic (N = 478)	69.12% (7.03)	1.11 (.47)	-.08 (.47)	.02 (.45)	$t(477) = 5.08$, $p < .001$
Study 5 Condition					
Control (N = 505)	67.24% (7.99)	1.00 (.52)	-.09 (.45)	.02 (.47)	$t(504) = 5.42$, $p < .001$
Bias Warning (N = 545)	67.98% (8.14)	1.04 (.52)	-.03 (.46)	-.02 (.46)	$t(544) = .62$, $p = .533$
Bias Warning + Delay (N = 445)	68.90% (8.24)	1.11 (.56)	-.06 (.47)	-.06 (.46)	$t(444) = -.08$, $p = .939$

Note. Criterion Bias = Within-subjects test comparing criterion values.

the rate at which errors of falsely rejecting qualified applicants occurred towards less versus more physically attractive people (Control = 53.6%, Delay = 52.9%). These results mirror the findings of Studies 2a-2c, where imposing shorter time windows resulted in more noise but again no changes in bias.

Study 3 data are consistent with the perspective that participants used the additional time to further parse the relevant academic criteria, resulting in fewer errors, but were unable or unwilling to use the delay in responding to counteract the bias in errors that favored more physically attractive applicants.

Whereas Studies 2a-3 manipulated the time available for participants to complete their judgments, Study 4 manipulated participants' motivation to engage in heuristic versus systematic processing of outcome-relevant information. Given that participants possess at least some ability to effectively parse the relevant criteria (as seen in above-chance levels of accuracy), motivation to process such information should moderate the level of noise in judgment. Heuristic thinking, which is associated with less cognitive scrutiny, should lead to more noise, while systematic thinking, which is associated with greater analysis of judgment-relevant information, should lead to less noise (Chen, Duckworth & Chaiken, 1999). Motivation to evaluate targets accurately may then provide a second avenue to reduce noise and lessen the impact of discrimination, and one that does so by directly manipulating individuals' mindsets rather than manipulating the judgment context (like in adding time pressure or requiring response delays).

Participants in Study 4 completed a JBT with no time pressure, but were instructed beforehand to either engage in more heuristic or systematic thinking. In addition, Study 4 investigated whether individual differences in thinking styles predict the degree of noise and bias in social judgment. Specifically, participants completed measures of faith in intuition and need

for cognition. These measures may add correlational evidence concerning how thinking styles are associated with noise and bias in social judgment.

Study 4

Method

Participants

We sought to collect at least 1200 eligible participants who completed the JBT across the three conditions. This sample size would provide greater than 80% power at detecting a small between-subjects effect of d = .20 between any two conditions. Exclusion rates were difficult to estimate, and the final sample was slightly larger. In total, 1574 participants (63.0% female, 72.0% White, MAge = 33.1, SD = 15.0) completed at least the JBT. See https://osf.io/kfas3/ for Study 4's pre-registration.

Procedure

Participants were randomly assigned to receive one of three instructions, followed by the JBT, measures of perceived performance, desired performance, and explicit preferences. They then completed individual differences measures of need for cognition and faith in intuition, followed by a measure of implicit evaluations of more versus less physically attractive people.

Academic JBT. Participants completed the same JBT as in Study 1a. Before beginning the JBT, participants were randomly assigned to the *Control*, *Heuristic* or *Systematic* condition. To increase similarity between conditions, Control participants were alerted to the fact that applicants would differ in ways other than their qualifications but were not told directly that applicants differed in physical attractiveness. A re-analysis of existing data found that this manipulation did not impact sensitivity or bias in criterion on the JBT relative to a condition receiving no additional instructions (Axt & Nosek, 2018). Specifically, Control participants read:

In addition to differing on their qualifications, applicants will differ in other ways. Prior research suggests that decision makers are easier on some types of applicants and tougher on other types of applicants.

To increase systematic thinking, participants in the Systematic condition were told to think hard about their judgments when completing the JBT. Specifically, participants in the Systematic condition read:

Prior research suggests that people may do a better job on this task if they put in more time to deliberate and think over their decisions. As a result, it is important that you think hard and slow down when making your decisions.

Participants completed a short writing task immediately before the test portion of the JBT where they listed two reasons why it was important to think hard and deliberate when making decisions.

To increase heuristic thinking, participants in the Heuristic condition were told to trust their first impressions and not to overthink their decisions when completing the JBT. Specifically, participants in the Heuristic condition read:

Prior research suggests that people may do a better job on this task if they trust their initial instincts and do not overthink their decisions. As a result, it is important that you 'go with your gut' and make your decisions more quickly.

Participants then listed two reasons why it was important to trust one's initial impressions when making decisions (see Appendix D for wording of both writing tasks).

Other measures. Participants completed the same measures of perceived performance, desired performance, explicit preferences and implicit associations as Study 3.

Participants also completed the 12-item Faith in Intuition scale (Epstein, Pacini, Denes-Raj, & Heier, 1996; α = .82, sample item: "I trust my initial feelings about people")[7] and the 18-item Need for Cognition scale (Cacioppo, Petty & Kao, 1984; α = .88, sample item: "I find satisfaction in deliberating hard and for long hours") in a randomized order. Both measures used a 1=Strongly disagree to 7 =Strongly agree response scale (Faith in Intuition M = 4.73, SD = .87; Need for Cognition M = 4.98, SD = .88).

Results

Using the same criteria as Study 3, 5.9% of participants were excluded from JBT analyses.

In an initial test of the effectiveness of the motivation manipulation, we compared average JBT reaction times (natural-log transformed) across conditions. Results suggested that our manipulation had the expected impact on time spent evaluating applicants. A one-way ANOVA on average reaction times found reliable differences across conditions, $F(2,1477) = 39.11$, $p < .001$, $\eta^2p = .050$, 95% C.I. [.03, .07]. Follow-up comparisons found that participants in the Heuristic condition had faster JBT responses than participants in the Control condition, $t(1000) = 3.61$, $p < .001$, $d = .23$, 95% C.I. [.10, .35], and the Systematic condition, $t(966) = 9.22$, $p < .001$, $d = .59$, 95% C.I. [.46, .72]. Participants in the Control condition in turn had faster reaction times than participants in the Systematic condition, $t(988) = 5.15$, $p < .001$, $d = .33$, 95% C.I. [.20, .45]. Transformed back into milliseconds, the Heuristic condition had an average reaction time of 1956 ms per judgment (SD =174), the Control condition an average of 2232 ms (SD = 184), and the Systematic condition an average of 2689 ms (SD = 169).

[7] In our pre-registration, we indicated we would compare the reliability of the full 12-item scale to a 9-item scale that excluded items about visual imagery (e.g., "I am good at visualizing things"). Reliability was higher for the 9-item version, so analyses use this shortened version.

Sensitivity and Bias in Decision-Making

A one-way ANOVA on sensitivity found reliable differences across conditions, $F(2,1371) = 15.75$, $p < .001$, $\eta^2p = .021$, 95% C.I. [.01, .04]. Follow-up comparisons found that participants in the Heuristic condition had lower sensitivity than participants in the Control condition, $t(1000) = 2.40$, $p = .016$, $d = .15$, 95% C.I. [.03, .28], and the Systematic condition, $t(966) = 5.73$, $p < .001$, $d = .37$, 95% C.I. [.24, .50]. Participants in the Control condition had lower sensitivity than participants in the Systematic condition, $t(988) = 3.21$, $p = .001$, $d = .20$, 95% C.I. [.08, .33]. Accuracy in all conditions was above chance (all t's > 45.32, all p's <.001).

All conditions exhibited bias, with lower criterion for more versus less physically attractive applicants (all t's > 5.08, all p's <.001, all d's > .23; see Table 4). A one-way ANOVA comparing the size of the criterion difference score found no reliable differences across conditions, $F(2,1477) = 1.00$, $p = .367$, $\eta^2p = .001$. Follow-up comparisons affirmed there were no reliable differences in bias between the Heuristic and the Control condition, $t(1000) = 0.98$, $p = .327$, $d = .06$, 95% C.I. [-.06, .19], or the Systematic condition, $t(966) = 1.37$, $p = .172$, $d = .09$, 95% C.I. [-.04, .21], as well as between the Control and Systematic condition, $t(988) = 0.41$, $p = .685$, $d = .03$, 95% C.I. [-.10, .15].

Associations with Need for Cognition and Faith in Intuition

Across conditions, greater sensitivity was weakly but reliably associated with greater need for cognition ($r = .172$, $p < .001$, 95% CI [.12, .21]) and lower faith in intuition, ($r = -.082$, $p = .002$, 95% CI [-.13, -.03]). Conversely, greater criterion bias was weakly but reliably associated with lower need for cognition ($r = -.090$, $p = .001$, 95% CI[-.14, -.04]) and greater

faith in intuition (r = .123, p < .001, 95% CI[.07, .17]). See online supplement for correlation matrix across all dependent variables.[8]

Discussion

Inducing systematic or heuristic thinking changed the amount of discrimination based on physical attractiveness. Participants who engaged in more heuristic thinking showed faster JBT responses and more noise (lower sensitivity). Participants who engaged in more systematic thinking showed slower JBT responses and less noise (higher sensitivity). In other words, the Study 4 manipulations impacted discrimination by making errors more or less likely (Heuristic = 21.4 errors, Control = 20.8 errors; Systematic = 19.8 errors). However, neither manipulation impacted bias, as all conditions showed comparable levels of a lower criterion for more versus less physically attractive applicants. That is, conditions did not vary in the rate at which errors of falsely admitting less qualified applicants occurred for more versus less physically attractive people (Heuristic = 53.8%; Control = 53.0%; Systematic = 53.7%), or in the rate at which errors of falsely rejecting more qualified applicants occurred for less versus more physically attractive people (Heuristic = 52.7%; Control = 53.3%; Systematic = 51.2%). Study 4 results are then similar to those of Studies 2a-3, which directly manipulated the time available for participants to respond.

Providing participants with greater motivation to process outcome-relevant information decreased noise, resulting in fewer errors, but had no impact on the proportion of those errors that favored one social group over another. That said, while the heuristic and systematic thinking manipulations led to changes in participants' reaction times and overall accuracy, Study 4 failed

[8] We also tested for interactions between experimental condition (with Control coded as the reference) and faith in intuition and need for cognition on overall sensitivity. There was no reliable interaction for faith in intuition (B =-.10, t = -.66, p = .507), or for need for cognition (B =.21, t = -1.42, p = .155).

to include a measure that allowed for a direct test of changes in participant motivation (e.g., a self-report item assessing desire to evaluate applicants accurately and based solely on their objective academic criteria). As a result, it's possible that the Study 4 manipulation impacted an outcome other than motivation. For instance, participants may have used the manipulation to create a self-imposed response deadline to either speed up in the heuristic condition or slow down in the systematic condition, a process that could have created changes in task sensitivity but not individual motivation *per se*. A goal of future work should be to include more straightforward measures of motivation when using this same manipulation, as well as to test whether other manipulations more closely tied to changes in motivation produce similar effects (e.g., incentivizing accurate performance).

Correlational analyses in Study 4 found weak but reliable associations between JBT sensitivity and faith in intuition as well as need for cognition, such that faith in intuition was negatively and need for cognition was positively associated with greater sensitivity. These small but reliable effects suggest that while the constructs of need for cognition and faith in intuition may be generally related to one's capacity or motivation to process outcome-relevant information in judgment, performance on this JBT specifically may be quite domain-specific (e.g., more related to faith in intuition about evaluating honor society applicants or in using physical attractiveness).

Perhaps surprisingly, bias (relative differences in criterion) was also reliably and weakly correlated with need for cognition and faith in intuition, such that faith in intuition was positively and need for cognition was negatively associated with greater bias. In our pre-registration, we anticipated the possibility that these individual difference measures would be associated with both noise and bias. For example, perhaps participants high in need for cognition may spend

more time parsing the relevant academic criteria (leading to a positive association with sensitivity) and also be less impacted by irrelevant social information when making errors (leading to a negative association with bias). These measures do not appear to assess processes distinctly associated with noise or bias in social judgment, an issue we return to in the General Discussion.

Studies 2a-4 provide strong evidence that the psychological processes dictating the degree of noise in social judgment differ from those dictating the degree of bias. Specifically, noise appears to be most tied to participants' ability and motivation to process outcome-relevant information, whereas these factors are less related or even unrelated in determining the level of bias in social judgment. Conversely, bias may be most tied to participants' ability or motivation to avoid the influence of irrelevant demographic information (e.g., giving participants greater ability to control bias in decision-making by alerting them to the demographic information likely to influence judgment). Study 5 tested this idea directly by investigating whether an intervention warning participants to the social dimension responsible for creating favoritism in evaluation, and asking them to avoid such favoritism, would impact bias, noise, or both.

Study 5

The interventions used in Studies 2a-4 impacted noise but had no effect on bias, which raises the question of what interventions impact bias. A separate line of research also using the JBT (Axt & Nosek, 2018) has identified several interventions that reduce bias (i.e., relative differences in criterion). Four interventions -- warning about the potential for biased judgment, committing to objective behavior beforehand, heightening accountability, or creating implementation intentions -- all reduced bias on the JBT (all p's $< .003$, all d's $> .17$; Axt &

Nosek, 2018, Study 1), provided that each intervention told participants that applicants would differ on a specific social dimension (here, physical attractiveness).

While the studies in Axt and Nosek (2018) were focused specifically on what is needed for interventions to reduce criterion bias, a re-analysis of the Study 1 data found that none of these interventions impacted noise relative to a Control condition (i.e., sensitivity; all p's > .112, all d's < .09). In these cases, interventions allowed participants to counteract favoritism towards more physically attractive applicants, but doing so led to overcorrection-- participants were now more stringent on all more physically attractive applicants and more lenient on all less physically attractive applicants, regardless of actual qualifications. The end result was a reduction in bias but no overall change in noise.

In Study 5, we sought to replicate and extend the findings of Axt and Nosek (2018) by using one intervention previously found to reduce criterion bias (alerting participants to the social dimension responsible for favoritism in judgment) and directly testing whether this intervention, unlike those used in Studies 3-4, impacted bias but not noise. These results would provide further support for the claim that different psychological processes impact the degree of bias versus noise in judgment.

In addition, we tested whether interventions can be combined to impact bias and noise simultaneously by including a condition where participants were both alerted to a tendency to favor more physically attractive people and had a required delay in responding. If effective, this intervention would demonstrate one means to increase accuracy and fairness simultaneously. However, it is possible that simply combining interventions will be ineffective if participants' attention becomes overly divided and the effectiveness of each individual intervention is reduced or eliminated (e.g., people can easily follow directions to rub their stomach or pat their heads, but

have a hard time completing both tasks simultaneously). It is also possible that, in a combined intervention, participants could attend to only one component of the intervention, leaving the other component ineffective. Finally, combining interventions could fundamentally alter how each individual intervention operates-- for instance, when alerted to a tendency to favor physically attractive people, participants may use the mandatory delay in responding to only further counteract the influence of attractiveness and not to better attend to the outcome-relevant criteria necessary for decreasing noise.

Method

Participants

Study 5 originally sought to collect at least 219 eligible participants per experimental condition. This sample size would provide 85% power for detecting a between-subjects effect of $d = .287$, which was the size of the reduction in criterion bias found previously with the same bias warning manipulation (see Axt & Nosek, 2018, Studies 1-3). Results from this initial sample were in the expected direction but inconclusive, so we pre-registered an additional data collection to double our sample size. Given multiple rounds of data analysis, primary analyses report *p*-augmented (Sagarin, Ambler & Lee, 2014), which is the inflated Type I Error rate that comes from completing multiple waves of data collection.

In total, 1601 participants (64.2% female, 66.8% White, *M*Age = 31.8, *SD* = 13.9) completed at least the JBT. See https://osf.io/bzwnh/ for Study 5's initial pre-registration and https://osf.io/rx4ey/ for updated analysis plan and results from the first round of data collection.

Procedure

Participants first completed the JBT, then measures of perceived performance, desired performance, explicit preferences and implicit associations.

Academic JBT. Participants completed the same JBT as in Study 1a, and were randomly assigned to a *Control*, *Bias Warning* or *Bias Warning + Delay* condition. Participants in the *Control* condition completed the task without additional instructions. Participants in the Bias Warning condition were alerted ahead of time to the possibility that they may favor physically attractive applicants over less attractive ones (Axt, Casola & Nosek, 2018; Axt & Nosek, 2018). Specifically, participants read:

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

Participants in the Bias Warning condition were also told immediately before the JBT to try and avoid favoring more over less physically attractive applicants. Finally, participants in the Bias Warning + Delay condition received the same intervention, but were also required to wait 4.5 seconds before responding on the JBT. Participants in the Bias Warning + Delay condition were told beforehand about the waiting period and completed a brief writing task where they listed two reasons why it was important to think hard about one's decisions.

Other measures. Participants completed the same measures of perceived performance, desired performance, explicit preferences, and implicit associations as in Study 4.

Results

Using the same criteria as Study 4, 6.6% of participants were excluded from JBT analyses. As in previous research (Axt, et al., 2018), a large majority of participants in all three conditions reported a motivation to not use physical attractiveness in their admissions judgments (Control: 89.3%, Bias Warning: 91.1%; Bias Warning + Delay: 88.2%).

Sensitivity and Bias in Decision-Making

A one-way ANOVA on overall sensitivity found reliable differences across conditions, $F(2,1492) = 5.73$, $p = .003$, $\eta^2p = .008$, 95% C.I. [.001, .018], *p*-augmented [.0506, .0508]. Follow-up comparisons found that participants in the Bias Warning + Delay condition had higher sensitivity than participants in the Control condition, $t(948) = 3.34$, $p = .001$, $d = .22$, 95% C.I. [.09, .35], *p*-augmented [.05004, .0502], and the Bias Warning condition, $t(988) = 2.12$, $p = .035$, $d = .14$, 95% C.I. [.01, .26], *p*-augmented [.066, .071]. The Control and Bias Warning conditions did not reliably differ in overall sensitivity, $t(1048) = 1.35$, $p = .177$, $d = .08$, 95% C.I. [-.04, .20]. Accuracy in each condition was above chance (all t's > 48.36, all p's < .001).

Participants in the Control condition showed bias, with lower criterion for more versus less physically attractive applicants ($d = .24$), whereas participants in the Bias Warning and Bias Warning + Delay condition showed no evidence of bias, meaning no reliable differences between criterion for more versus less physically attractive applicants (all t's < .62, all p's < .533; see Table 4). A one-way ANOVA on criterion bias difference scores found reliable differences across conditions, $F(2,1492) = 8.81$, $p < .001$, $\eta^2p = .012$, 95% C.I. [.003, .024], *p*-augmented [.050005, .05001]. Follow-up comparisons found that participants in the Control condition had greater bias than participants in the Bias Warning condition, $t(1048) = 3.55$, $p < .001$, $d = .22$, 95% C.I. [.10, .34], *p*-augmented [.05, .05005], and the Bias Warning + Delay condition, $t(948) = 3.68$, $p < .001$, $d = .24$, 95% C.I. [.11, .37], *p*-augmented [.05, .05003]. The Bias Warning and Bias Warning + Delay conditions did not reliably differ, $t(988) = 0.46$, $p = .645$, $d = .03$, 95% C.I. [-.10, .15].

Discussion

Whereas the interventions used in Studies 2-4 impacted noise but not bias, the bias warning intervention used in Study 5 had the reverse effect of impacting bias but not noise. Both of the experimental interventions used in Study 5 reduced discrimination, but did so through different routes. Warning participants of a tendency to favor more physically attractive people reduced the rate at which errors of falsely accepting less qualified applicants occurred towards more versus less attractive people (Control = 52.9%, Bias Warning = 50.1%, Bias Warning + Delay = 50.5%) and the rate at which errors of falsely rejecting more qualified applicants occurred towards less versus more attractive people (Control = 52.4%, Bias Warning = 50.4%, Bias Warning + Delay = 49.4%). However, only requiring a response delay led to reliably fewer errors (Control = 21.0 errors, Bias Warning = 20.5 errors, Bias Warning + Delay = 20.0 errors).

The effectiveness of this bias warning manipulation replicates past work (Axt & Nosek, 2018; Axt, Casola & Nosek, 2018), but more specifically illustrates how alerting participants beforehand to a tendency to favor more physically attractive people reduces discrimination. The bias warning intervention reduced discrimination not by changing how many errors were made but instead by changing the distribution of errors such that more and less physically attractive people were equally likely to receive beneficial versus detrimental treatment. Moreover, Study 5 extends these prior studies by revealing how participants exposed to this same bias warning manipulation that were also required to wait 4.5 seconds before responding showed both reduced bias and reduced noise (i.e., greater sensitivity). These results confirm that it is possible for a "single" intervention to simultaneously lessen noise and bias, as reductions in bias did not come at the price of overcorrection.

It is notable that, in Study 5, interventions warning about favoritism towards physically attractive people were strong enough to eliminate rather than merely reduce the criterion bias in evaluation (and therefore eliminate discrimination based on physical attractiveness). These data, coupled with other uses of this same bias warning manipulation that also eliminated criterion biases favoring more attractive applicants on the JBT (Axt & Nosek, 2018; Studies 1 and 3), provide promising evidence of a particularly impactful intervention for reducing discrimination in social judgment. However, it is worth pointing out that the straightforward bias warning manipulation used in Study 5 may not produce comparable effects in more socially sensitive dimensions like race or gender, where participants could either dismiss the possibility that they may be biased or exert even higher levels of motivation to be unbiased. Though other work has used a similar manipulation to reduce political ingroup biases in judgment (Axt, Casola & Nosek, 2018), it will be important for future studies to test whether such interventions extend to other social dimensions. Finally, it remains unclear whether the relatively minimal bias warning used here would effectively reduce discrimination outside of a judgment context immediately following the manipulation; for more durable and generalizable changes in behavior, it will likely be necessary to include stronger and more immersive interventions (Bezrukova, Spell, Perry, & Jehn, 2016, Forscher et al., 2017).

General Discussion

We investigated two distinct outcomes in determining the magnitude of discrimination in social judgment: noise-- the overall number of errors made in evaluation-- and bias-- the degree to which errors favor one group over another. Using a judgment task that had objectively correct or incorrect answers, Studies 1a-1b found that the presence of non-diagnostic social information not only created bias in the distribution of errors but also increased the total number of errors

made, revealing how such social information is actively incorporated into judgment at the expense of greater use of more outcome-relevant criteria. These studies provide the first experimental evidence that removing irrelevant social information from decision-making not only eliminates the possibility for bias but also improves the accuracy of evaluation, lending further support for the practice of "blinding" applications, resumes, or other means of evaluation when possible and appropriate (Gouldin & Rouse, 2000; c.f. Doleac & Hansen, 2018).

In Studies 2a-5, the degree of bias and noise in judgment was differentially impacted by various interventions. Just as motivation and ability are key determinants of the strength of attitude-behavior correspondence (Fazio & Olson, 2003), they may also dictate the magnitude of noise and bias in social judgment. Specifically, manipulations that altered participants' ability or motivation to process outcome-relevant information changed the degree of noise but not bias. Conversely, a manipulation that potentially increased participants ability and motivation to monitor the impact of irrelevant social information, by warning them of an influence on judgment that they may have otherwise failed to noticed or taken seriously, reduced bias but not noise. Finally, an intervention that both increased the time available to process outcome-relevant information and alerted decision-makers to the influence of irrelevant social information reduced both bias and noise, revealing that interventions can change both outcomes simultaneously.

This work provides additional evidence that warning participants of their potential biases can occasionally reduce those same biases in judgment (Golding, Fowler, Long, & Latta, 1990; Pope, Price & Wolfers, 2016; Pronin & Krugler, 2006; Schul, 1993). In addition, the effectiveness of the bias warning manipulation used in Study 5 lends support to various diversity interventions that have all alerted participants to their own prejudices in efforts to reduce biased behavior or increase egalitarian motivations (e.g., Carnes et al., 2012; Devine et al., 2017;

Forscher et al., 2017; Pietri et al., 2016). The present results suggest that part of the power of these interventions is in making errors more evenly distributed in judgment contexts (i.e., reducing bias). Finally, our results align with a more recent investigation finding that discrimination can be lessened by simply increasing the accuracy of the decision-making process (Chang & Cikara, 2018). At the same time, the present work extends these prior investigations by illustrating how warnings about potential bias can lead to reductions in the relative likelihood of certain groups receiving favorable treatment without influencing the *amount* of people receiving unfair treatment. Similarly, increasing the accuracy of the decision-making can reduce the amount of people receiving unfair treatment yet preserve the relative degree of favoritism for some groups over others when errors do occur.

Using a relatively novel measure of social judgment allowed for a more nuanced investigation into the impact of these previously strategies to reduce bias or discrimination. Our results highlights the need for additional measures of discrimination that lend themselves to a signal detection analysis, where behavior can be evaluated in terms of bias and noise. Many prior investigations of interventions seeking to reduce discrimination have used outcomes that leave it ambiguous as to whether such interventions are effective via changes in noise and/or bias (e.g., Bohnet, van Geen, & Bazerman, 2015; Uhlmann & Cohen, 2005). As many existing measures of discriminatory behavior lack validation and suffer from low reliability (Carlsson & Agerström, 2016), greater empirical and theoretical progress on research into discrimination will come from the use of validated measures that allow for more nuanced analyses like those presented here (Greenwald, 2012).

Bias and Noise as Distinct but Related Components of Social Judgment

Throughout this work, we treated bias and noise as conceptually independent outcomes in determining the magnitude of discrimination in social judgment. Bias and noise are analytically independent, in that knowing the value of one tells you virtually nothing about the other (Green & Swets, 1966), but practically there are many reasons to believe that there is some relationship between the two. It is certainly plausible that there is a correlation between the likelihood of making errors in general and the likelihood of those errors revealing more favorable treatment for some groups over others (e.g., participants more motivated to favor a certain group may also be less motivated to attend to the relevant qualifications when completing their judgments).

Indeed, bias and noise may share some overlapping causes. A meta-analysis across studies found a small, negative correlation between the criterion bias difference score and overall sensitivity (r = -.131, 95% CI [-.168, -.094]; see online supplement for results from each study). This negative correlation indicates that participants who were more prone to making errors on the JBT or the FPST (i.e., lower sensitivity and higher noise) were also more likely to have those errors disproportionately favor one group over another (i.e., greater differences in criterion between social groups), though this relationship was relatively weak.

Study 4 results provide additional evidence for a conceptual link between bias and noise, and shed some light on the psychological constructs that may be dually associated with the two outcomes. Greater need for cognition was associated with decreased criterion bias and increased sensitivity, while greater faith in intuition was associated with increased criterion bias and decreased sensitivity. These data align with prior work that finds a consistent relationship between thinking styles and psychological processes like stereotyping (e.g., Kearney, Gebert & Voelpel, 2009; Schaller, Boyd, Yohannes & O'Brien, 1995; Trent & King, 2013).

At the same time, Studies 2b-2c provide preliminary evidence that more social motivations, such as the desire to avoid prejudice, are also jointly related to the degree of bias and noise in judgment. Participants completed an adapted versions of the internal motivation to avoid prejudice (IMS; Plant and Devine, 1999) scale as well as an attractiveness JBT. Across studies, greater IMS was weakly but negatively related with criterion bias (r = -.085, 95% CI [-.128, -.042]) and positively related with overall sensitivity (r = .122, 95% CI [.080, .165]). Though effects were again small, they extend past studies investigating behavioral outcomes associated with motivation to control prejudiced reactions (Amodio, Harmon-Jones & Devine, 2003; Payne, 2005; Gonsalkorale, Sherman, Allen, Klauer & Amodio, 2011).

Measures of thinking styles did not solely predict outcomes more associated with attending to relevant information (i.e., sensitivity), and measures of motivation to control biased responses did not solely predict outcomes more associated with ignoring irrelevant information (i.e., criterion bias). Rather, bias and noise were at least weakly associated with individual differences in both social and cognitive processes. Future research on this topic should continue to explore how and when individual differences in desires to be accurate versus desires to control one's social biases independently and jointly contribute to the magnitude of discrimination. In addition, individual differences in other psychological processes may also be related to both bias and noise, such as executive function (Ito et al., 2015; Payne, 2005) or chronic egalitarianism (Moskowitz & Li, 2011).

Overall, these results suggest that bias and noise are distinct but related. The manipulations within these studies consistently affected one without the other, and the correlational relationship between the two was consistently weak. However, the apparent

association between the two is an encouraging sign for those seeking to design interventions that impact both bias and noise simultaneously.

Implications for Automaticity and Cognitive Resources in Bias

Some of the more striking data from the present work come from Studies 2a-2c, where added time pressure strongly increased the level of noise in social judgment but had no reliable impact on bias in criterion. These results appear to conflict with prior perspectives that suggest processes like stereotypes and prejudice are partly automatic and are therefore exacerbated when cognitive resources are constrained. Indeed, several prior studies have found that social biases--either in judgment, attitudes, or perceptions-- are heightened when cognitive resources are low. For example, participants asked to remember an eight-digit number during a learning task then showed greater retention of stereotype-consistent than stereotype-inconsistent information relative to control participants (Sherman, Lee, Bessenoff & Frost, 1998). Participants receiving the same or similar manipulations also showed increased activation of race-related stereotypes on a word-fragment completion task (Gilbert & Hixon, 1991), or exhibited less contextual sensitivity in a measure of person perception (Gilbert, Pelham & Krull, 1988). Given these earlier studies, one possible expectation is that participants completing the FPST or JBT in contexts where cognitive resources are limited, such as in the deadline conditions in Studies 2a-2c, should show greater activation of relevant stereotypes concerning race or attractiveness, which would lead to greater favoritism in judgment and increased bias in criterion.

At the same time, studies using outcomes and manipulations similar to those in the current work have failed to find that greater depletion or fewer cognitive resources impacts bias in criterion. For instance, completing a prolonged Stroop task (to induce feelings of depletion) produced more errors in a Weapons Identification Task but no changes on the level of bias in

criterion (Govorun & Payne, 2006). In a more striking demonstration, Correll, Wittenbrink, Crawford and Sadler (2015) used a modified FPST where participants provided two responses on each trial. The first response had to occur within the 630 milliseconds that the target appeared onscreen. After the target disappeared, participants provided a second response where they could take as long as needed. Follow-up responses produced fewer errors and greater sensitivity than the initial speeded responses, but relative biases in criterion (i.e., a lower threshold to indicate shooting for Black versus White targets) did not reliably differ between the two response formats. These results closely mirror those of Studies 2a-2c in that increasing capacity to process outcome-relevant information did not translate into reduced biases in criterion.

Some of the discrepancies between the current work and past research suggesting that bias is partly automatic may be attributed to an inability for past social cognition research to distinguish between noise and bias in group-based disparities. Our approach separates the two by considering bias to be a *proportional* outcome-- the relative likelihood of committing certain types of errors for some groups versus others. One consequence of using a proportional outcome is that some manipulations, like added time pressure, can raise the absolute value of these relative differences in the types of errors committed for some groups over others, but signal detection analyses will find no changes in bias if those errors accrue at the same rate as when judgments are made under less or no time pressure.

For example, the time pressure manipulations of Studies 2b-2c resulted in a larger number of more attractive applicants receiving beneficial treatment and less attractive applicants receiving detrimental treatment, but this effect could be attributed to a general increase in errors rather than a greater rate of favoring more over less attractive applicants when errors did occur. This distinction is obscured in many prior studies that used outcomes lacking objectively correct

answers to distinguish bias from noise (e.g., Gilbert & Hixon, 1991; Gilbert, Pelham & Krull, 1988). In studies of social judgment that used outcomes lacking a correct response, conceptions of "bias" were closer to our use of discrimination-- favoritism in treatment for one social group over another (e.g., Bodenhausen, Kramer & Susser, 1994; Norton, Vandello & Darley, 2004). The signal detection framework applied here and in recent studies help clarify the role of cognitive resources in social judgment and the degree to which social judgment is impacted by automatic processes. Cognitive resources do not appear to moderate the relative strength at which social biases impact judgments when errors are made; rather, fewer cognitive resources leads to more errors in general, and discrimination is increased when the same degree of bias operates over a larger number of errors.

Preferences for Interventions that Reduce Bias versus Noise

One practical reaction to the distinction between bias versus noise may be whether interventions that lessen discrimination by reducing bias are any "better" or "worse" than interventions that lessen discrimination by reducing noise. One approach-- reducing noise-- minimizes the amount of people who receive unfair treatment but preserves the unequal dispersion of that unfair treatment across demographic groups. The other approach-- reducing bias-- preserves the amount of people who receive unfair treatment but reduces the degree to which that unfair treatment is more likely to impact one social group versus another.

We think the more effective approach depends on the situation. For instance, in outcomes where criteria are more straightforward and errors are highly costly (e.g., falsely shooting people who pose no threat of violence), improving sensitivity (i.e., lessening the total number of people who are unnecessarily shot) may be more beneficial than reducing bias based on targets' demographic information. Conversely, in outcomes where the criteria associated with improved

accuracy are hard to determine and errors are relatively less costly, it may be more practical to lessen discrimination by targeting a reduction in bias. For instance, it is difficult to establish objective criteria for what constitutes an effective instructor; individuals may rightly weight various criteria differently-- e.g., clarity of presentation, speed of receiving feedback, transparency in evaluation. In such cases, creating interventions that improve sensitivity (i.e., reducing the amount of worthy instructors who receive negative evaluations and the amount of unworthy instructors who receive positive evaluations), may be quite difficult. Instead, efforts to reduce discrimination in these contexts, like in addressing gender disparities in teaching evaluations (Mengel, Sauermann & Zolitz, 2018), may be better served by focusing on reducing bias, such as in directly warning students to avoid using gender in their evaluations.

In many cases, it is not necessary to choose between the two approaches at all. It is often most effective to pursue both approaches simultaneously, as we have shown in Study 5. For example, hiring managers could enforce standardized criteria in evaluating job candidates to reduce noise and provide warnings to avoid using demographic information to reduce bias. Employing both approaches in concert may lead to greater reductions in discrimination than using either approach in isolation.

Regardless of effectiveness, people may have preferences for what approaches should be used in everyday life. From an initial investigation, we suspect that lay preferences for how to reduce discrimination will be context-dependent and shift based on factors like the type of decisions being made and the social groups receiving the unfair treatment. We asked a sample of Americans (N = 1519) to read one of three vignettes that detailed various forms of discrimination. Each vignette described an organization seeking to address discrimination-- a company trying to reduce favoritism towards more physically attractive candidates in hiring, a

university math department trying to reduce preference for male applicants in graduate admissions, or a police department trying to reduce racial disparities in shooting unarmed suspects. Participants were told that the organization needed to choose between two training sessions that would reduce this discrimination. One training session would lessen bias-- no overall change in the number of errors made, but a more equal distribution of errors across social groups. The other would essentially lessen noise-- reduce the overall number of errors made, but some groups would still be more likely to receive unequal treatment. Participants then chose which training program should be implemented.

We found that preferences for training programs depended on the context. 52.1% of participants chose the bias-lessening program over the noise-lessening program when the goal was to reduce gender-based discrimination in admissions. However, 60.9% chose the noise-lessening program instead when the goal was to reduce race-based discrimination in police shootings, $\chi^2(1) = 16.96$, $p < .001$ (see online supplement for full methods and analyses). Though effects are relatively small, these data suggest that preferences for reducing discrimination via changes in bias or noise may vary based on the type or severity of the discrimination. Future work in this area will benefit from further exploring what factors drive preferences for changes in bias or noise as a means for reducing discrimination.

Future Directions and Generality

Subsequent research into the role of bias and noise in discrimination should look to extend and clarify many of the claims made here. For one, it will be important to identify boundary conditions where increased systematic thinking or ability to regulate bias are ineffective. There are likely conditions that need to be met beforehand in order for such interventions to be effective. One candidate is task difficulty. In tasks where it is challenging to

determine the correct answer, interventions that increase systematic thinking or heighten ability to control socially biased responding may be ineffective, as participants will lack the capacity to exert much mental control over responses (Pehman & Neter, 1995; Wilson & Brekke, 1994). The JBT may offer one means of testing this proposition directly, as task difficulty can be manipulated by altering the ease at which participants can identify the appropriate responses (i.e., by shrinking the relative gap in qualifications between more and less qualified applicants).

In addition, it will be of practical and theoretical value to identify additional interventions that reduce bias and noise simultaneously. Studies 1a-1b identified blinding as a strategy for doing so. Removing irrelevant social information eliminates the possibility of bias and reduces noise, but blinding is not feasible in many evaluation contexts (e.g., in-person interviews). Another possibility is using a version of implementation intentions (Gollwitzer, 1999) that simultaneously directs attention towards the outcome-relevant criteria and away from the irrelevant social information (e.g., through having participants adopt the strategy of "When I see an application, I will ignore physical attractiveness and attend to the academic criteria"). Finding new approaches for reducing both bias and noise simultaneously will shed light on the mechanisms that give rise to these two components of social judgment.

Finally, though this work used multiple outcome measures and social groups as targets, samples were limited to a single source. Participants in all studies came from a volunteer website with a stated mission of investigating biases in attitudes and judgment. Though our large samples allowed for a relatively wide range of participants in terms of demographic characteristics like age, race, and political orientation, they are not representative of any definable population. It is possible that more representative samples would show differing effects from those presented here, though we cannot at the moment identify a plausible reason to expect this lack of

generalizability. Regardless, investigating the generality of this work to both other populations and other forms of discrimination will be a priority of future work.

Conclusions

Progress on reducing discrimination will come from greater clarity into the processes that give rise to discriminatory behavior and the interventions that alter these processes. This work identifies two distinct but related components of discrimination-- noise and bias-- and reveals how each are differentially impacted by various interventions and rely on distinct psychological processes. Future efforts to reduce discrimination should consider the relative ease and effectiveness of targeting changes in bias, noise, or both.

References

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of Personality and Social Psychology*, *84*(4), 738-753.

Ariely, D., & Zakay, D. (2001). A timely account of the role of duration in decision making. *Acta Psychologica*, *108*(2), 187-207.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological bulletin*, *110*(3), 486-498.

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

Axt, J.R., Casola, G.M, & Nosek, B.A. (2018). Reducing social judgment biases may require identifying the potential source of bias. *Personality and Social Psychology Bulletin*. Advance online publication: doi: 10.1177/0146167218814003.

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

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

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

Ayres, I., & Siegelman, P. (1995). Race and gender discrimination in bargaining for a new car. *American Economic Review*, *85*(3), 304-321.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991-1013.

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016). A meta-analytical integration of over 40 years of research on diversity training evaluation. *Psychological Bulletin*, *142*(11), 1227-1274.

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

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

Burns, M. D., Monteith, M. J., & Parker, L. R. (2017). Training away bias: The differential effects of counterstereotype training and self-regulation on stereotype activation and application. *Journal of Experimental Social Psychology*, *73*, 97-110.

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306-307.

Castellan Jr, N. J. (1974). The effect of different types of feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance*, *11*(1), 44-64.

Chang, L. W., & Cikara, M. (2018). Social Decoys: Leveraging choice architecture to alter social preferences. *Journal of Personality and Social Psychology*, *115*(2), 206-223.

Chen, S., Duckworth, K., & Chaiken, S. (1999). Motivated heuristic and systematic processing. *Psychological Inquiry*, *10*(1), 44-49.

Cicchetti, D. V. (1993). The reliability of peer review for manuscript and grant submissions: "It's like déjà vu all over again!". *Behavioral and Brain Sciences*, *16*(2), 401-403.

Cobb, J. (2016). The matter of black lives. *The New Yorker*, *14*, 33-40.

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.

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology*, *108*(2), 219-233.

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

Dhami, M. K. (2003). Psychological models of professional decision making. *Psychological Science*, *14*(2), 175-180.

Dhami, M. K., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making*, *14*(2), 141-168.

Dhami, M. K., & Harries, C. (2001). Fast and frugal versus regression models of human judgement. *Thinking & Reasoning*, *7*(1), 5-27.

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

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

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, *71*(2), 390-405.

EU-MIDIS. 2009. *European Union Minorities and Discrimination Survey.* Retrieved from http://fra.europa.eu/en/publication/2017/eumidis-ii-main-results.

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

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297-327.

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

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., et al. (under review). A meta-analysis of change in implicit bias. Manuscript submitted for publication.

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

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, *60*(4), 509-517.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, *54*(5), 733-740.

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, ethnicity, and NIH research awards. *Science*, *333*(6045), 1015-1019.

Gonsalkorale, K., Sherman, J. W., Allen, T. J., Klauer, K. C., & Amodio, D. M. (2011). Accounting for successful control of implicit racial bias: The roles of association activation, response monitoring, and overcoming bias. *Personality and Social Psychology Bulletin*, *37*(11), 1534-1545.

Govorun, O., & Payne, B. K. (2006). Ego—depletion and prejudice: Separating automatic and controlled components. *Social Cognition*, *24*(2), 111-136.

Green, T. K. (2003). Discrimination in workplace dynamics: Toward a structural account of disparate treatment theory. *Harvard. Civil Rights-Civil Liberties Law Review*, *38*, 91-157.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psycho-physics*. New York: Wiley. (Reprinted 1974, Huntington, NY: Krieger).

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

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197-216.

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

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology, 108*(2), 187-218.

Keinan, G. (1987). Decision making under stress: Scanning of alternatives under controllable and uncontrollable threats. *Journal of Personality and Social Psychology, 52*(3), 639-644.

Kruglanski, A.W. & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97-109.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480-498.

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review, 118*(2), 316-338.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*(4), 1122-1135.

Ma-Kellams, C., & Lerner, J. (2016). Trust your gut or think carefully? Examining whether an intuitive, versus a systematic, mode of thought produces greater empathic accuracy. *Journal of Personality and Social Psychology, 111*(5), 674-685.

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology, 52*(6), 352-361.

McDaniel, L. S. (1990). The effects of time pressure and audit program structure on audit performance. *Journal of Accounting Research*, 267-285.

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

Mengel, F., Sauermann, J., & Zölitz, U. (in press). Gender bias in teaching evaluations. *Journal of the European Economic Association*. Advance online publication.

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

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

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

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

Moulton, C. A., Regehr, G., Mylopoulos, M., & MacRae, H. M. (2007). Slowing down when you should: a new model of expert judgment. *Academic Medicine, 82*(10), S109-S116.

Neumark, D., Bank, R. J., & Van Nort, K. D. (1996). Sex discrimination in restaurant hiring: An audit study. *The Quarterly Journal of Economics*, *111*(3), 915-941.

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of personality and social psychology*, *87*(6), 817-831.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565-584.

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., & Smyth, F. L. (2007). A multitrait-multimethod validation of the implicit association test. *Experimental Psychology*, *54*(1), 14-29.

Park, J., & Banaji, M. R. (2000). Mood and heuristics: The influence of happy and sad states on sensitivity and bias in stereotyping. *Journal of Personality and Social Psychology*, *78*(6), 1005-1023.

Payne, B. K. (2005). Conceptualizing control in social cognition: how executive functioning modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology*, *89*(4), 488-503.

Pelham, B. W., & Neter, E. (1995). The effect of motivation of judgment depends on the difficulty of the judgment. *Journal of Personality and Social Psychology*, *68*(4), 581-594.

Pietri, E. S., Moss-Racusin, C. A., Dovidio, J. F., Guha, D., Roussos, G., Brescoll, V. L., & Handelsman, J. (2017). Using video to increase gender bias literacy toward women in science. *Psychology of Women Quarterly*, *41*(2), 175-196.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811-832.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*, e3677.

Ross, S. L., & Turner, M. A. (2005). Housing discrimination in metropolitan America: Explaining changes between 1989 and 2000. *Social Problems*, *52*(2), 152-180.

Ross, S. L., & Yinger, J. (2002). *The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement*. MIT Press.

Rothstein, H. G. (1986). The effects of time pressure on judgment in multiple cue probability learning. *Organizational Behavior and Human Decision Processes*, *37*(1), 83-92.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*(3), 293-304.

Sattler, D. N., McKnight, P. E., Naney, L., & Mathis, R. (2015). Grant peer review: improving inter-rater reliability with training. *PLoS One*, *10*(6), e0130450.

Sherman, J. W., Lee, A. Y., Bessenoff, G. R., & Frost, L. A. (1998). Stereotype efficiency reconsidered: Encoding flexibility under cognitive load. *Journal of Personality and Social Psychology*, *75*(3), 589-606.

Sommers, S. R., & Kassin, S. M. (2001). On the many impacts of inadmissible testimony: Selective compliance, need for cognition, and the overcorrection bias. *Personality and Social Psychology Bulletin*, *27*(10), 1368-1377.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology*, *56*(4), 283-294.

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, *34*(10), 1332-1345.

Troutman, C. M., & Shanteau, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance*, *19*(1), 43-55.

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, *16*(6), 474-480.

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

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

Xu, Q., & Shrout, P. E. (2018). Accuracy and bias in perception of distress level and distress change among same-sex college student roommate dyads. *Personality and Social Psychology Bulletin*, *44*(6), 899-913.

Appendix A

Criteria values for more and less qualified applicants. Applicants are scored by placing each criterion on a four-point scale. GPA's are already on this scale, and interview scores, which are displayed as out of 100 possible points are divided by 25. Rec. Letters are scored such that Poor = 1, Fair = 2, Good = 3, and Excellent = 4. Less qualified applicants have criteria that sum to a total score of 14, while more qualified applicants have criteria that sum to 14.

Less Qualified Applicants

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

3.1	3.7	Good	80
3.5	3.7	Good	70

More Qualified Applicants

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

Appendix B

Participants in Study 2a completed a four-block, danger-focal Brief Implicit Association Test (BIAT; Sriram & Greenwald, 2009), measuring the strength of the association between the concepts 'Danger' and 'Safe' and the categories 'White people' and 'Black people'. BIAT responses were scored by the *D* algorithm (Nosek, Bar-Anan, Sriram, Axt, & Greenwald, 2014), such that more positive scores reflected a stronger association between Black people and danger and White people and safety.

Each BIAT contained four blocks. In the first block (20 trials), participants pressed the "I" key for all Dangerous words (Dangerous, Risky, Alarming, Threatening) and gray-scale images of White people (cropped to only show the face, two male and two female), and the "E" key for "any other images and words." These other stimuli were Safety words (Safe, Secure, Harmless, Protected) and gray-scale images of Black people (cropped to only show the face, two male and two female). The second block (20 trials) had the same design, but the "I" key was for Safety words and images of Black people, and the "E" key for "any other images or words".

The third and fourth blocks repeated the first and second blocks, respectively. Danger was always the focal category and assigned to the "I" key, but participants were randomly assigned either to an order where the first and third blocks paired Danger words with White faces, or to an order where the first and third blocks paired Danger words with Black faces. Participants were excluded from analyses involving the BIAT if more than 10% of trial responses were faster than 300 milliseconds (4.9% of participants with BIAT scores; Nosek et al., 2014).

Appendix C

Study 3 writing task instructions.

Delay Condition

Again, it is critical that you think hard about each decision you make.

In the box below, please write out two reasons for why **careful deliberation** is important for deciding who to admit to an honor society.

Appendix D

Study 4 writing task instructions.

Heuristic Thinking Condition

Again, it is critical that you do not overthink your decisions and go with your "gut" response when making decisions.

In the box below, please write out two reasons for why **trusting your initial impressions** is important for deciding who to admit to the honor society.

Systematic Thinking Condition

Again, it is critical that you think hard about your decisions and deliberate about each of your decisions.

In the box below, please write out two reasons for why **thinking hard about your decisions** is important for judging who to admit to the honor society.