Journal of Experimental Psychology: Applied

Less Biased yet More Defensive: The Impact of Control Processes

Joseph A. Vitriol, Brian A. O’Shea, and Jimmy Calanchini

Online First Publication, March 9, 2023. https://dx.doi.org/10.1037/xap0000468

Journal of Experimental Psychology: Applied

© 2023 American Psychological Association
ISSN: 1076-898X

https://doi.org/10.1037/xap0000468

Less Biased yet More Defensive: The Impact of Control Processes

Joseph A. Vitriol[1], Brian A. O'Shea[1], and Jimmy Calanchini[2]
[1] Department of Psychology, Harvard University
[2] Department of Psychology, University of California, Riverside

Educational and training programs designed to reduce racial bias often focus on increasing people's awareness of psychological sources of their biases. However, when people learn about their biases, they often respond defensively, which can undermine the effectiveness of antibias interventions and the success of prejudice regulation. Using process (Quad) modeling, we provide one of the first investigations of the relationships between (a) controlled and automatic cognitive processes that underpin performance on the Implicit Association Test and (b) defensive reactions to unflattering implicit racial bias feedback. In two correlational samples (one preregistered; $N = 8{,}000$) and one experiment in which the provision of bias feedback was manipulated ($N = 547$), we find racially biased associations and some control over these associations among White people. Nonetheless, more defensiveness to bias feedback consistently predicted weaker ability to control biased associations. We also find correlational evidence that lower levels of biased associations predict more defensiveness, but did not replicate this observation in the experimental study. These results are critical for theories of implicit attitudes, models of prejudice regulation, and strategies for antibias interventions.

Public Significance Statement

We illuminate the psychological obstacles that undermine the kind of bias awareness that aids successful prejudice regulation and antibias interventions. Specifically, we answered an unaddressed question of theoretical and practical importance: How are biased associations and control processes related to defensive responding among those who receive unflattering implicit racial bias feedback? We find partial evidence that defensiveness is related to lower (vs. higher) levels of racially biased associations among White people. This finding joins recent research in challenging traditional assumptions about the unconscious nature of the Implicit Association Test. Additionally, we find that defensiveness was related to less control over biased responses. Thus, increasing perceived efficacy in controlling and regulating bias may be an effective approach to bias education and training. By reducing defensiveness, such an approach can promote bias awareness and therefore motivate increased commitment to fairness.

Keywords: implicit bias, diversity, prejudice and stereotyping, organizational behavioral, management science

Supplemental materials: https://doi.org/10.1037/xap0000468.supp

Educational and training programs designed to reduce racial bias often focus on increasing people's awareness of psychological sources of their biases (Carter et al., 2020; Moskowitz & Vitriol, 2021). Implicit attitudes, for example, can influence behavior independent of explicit beliefs (Kurdi et al., 2019). When people are not aware of how their implicit attitudes relate to bias, they are less able to achieve their egalitarian goals (Moskowitz, 2010). By highlighting a discrepancy between behavior and egalitarian norms,

Joseph A. Vitriol https://orcid.org/0000-0002-2715-6123

Joseph A. Vitriol is now at the Department of Management, Lehigh University College of Business.

Brian A. O'Shea is now at the School of Psychology, University of Nottingham.

Joseph A. Vitriol played lead role in conceptualization, writing of original draft and writing of review and editing, supporting role in data curation and equal role in formal analysis and methodology. Brian A. O'Shea played supporting role in conceptualization, writing of original draft and writing of review and editing and equal role in data curation, formal analysis and methodology. Jimmy Calanchini played lead role in formal analysis, supporting role in conceptualization, writing of original draft and writing of review and editing and equal role in data curation and methodology.

Preregistered procedure, hypotheses, and plan for analysis of confirmatory sample in Study 1 are available at: https://osf.io/9yjqc/?view_only=6b5e72c9456447c1ad39ed25f494d04c.

Data and data syntax for Studies 1 and 2 are available at https://osf.io/q264d/?view_only=ed84c0c937604b3591ef98f33077f500.

Codebook for 2013 Race Implicit Association Test from Project Implicit is available at https://osf.io/m8uxt.

Correspondence concerning this article should be addressed to Joseph A. Vitriol, Department of Management, Lehigh University College of Business, 621 Taylor Street, Bethlehem, PA 18015, United States. Email: joevitriol@gmail.com

awareness of bias can trigger the self-regulation of prejudiced responding (Monteith, 1993).

For example, Monteith et al. (2001) find that when low-prejudice people discover a discrepancy between their behavior and values, they experience negative affective arousal, which can motivate increased commitment to egalitarian goals and, consequently, increased control over prejudiced responding. Moskowitz et al. (1999) similarly demonstrate how prejudice-related discrepancies lead egalitarians to inhibit stereotypes and focus their attention on nonprejudiced goals (Moskowitz & Li, 2011). Awareness of bias is the first step toward bias reduction and control (Burns et al., 2017; Vitriol & Moskowitz, 2021). Providing people with information about their implicit attitudes is therefore a common and effective component of bias reduction initiatives (Axt et al., 2019; Hillard et al., 2013; Parker et al., 2018; Perry et al., 2015; Régner et al., 2019).

How people respond to this information is critical to understanding the self-regulation of bias and improving the effectiveness of antibias interventions. In this article, we report the results of one of the first investigations of the cognitive processes that underpin performance on the Implicit Association Test (IAT; Greenwald et al., 1998) and reactions to feedback about implicit racial bias.

Defensive Reactions to Implicit Bias Feedback

While increasing awareness of bias is necessary (although insufficient) for bias education and prejudice regulation, providing accurate information about bias can, unfortunately, threaten people's cherished self-views and normatively desirable self-presentations (Moskowitz & Vitriol, 2021; Vitriol & Moskowitz, 2021). Learning that one has acted or has the potential to act with bias can motivate sincere commitment to egalitarian behavior and, over time, an improved ability to regulate the expression of prejudice (Monteith et al., 2001). But such feedback can also be experienced as a threat to one's values and social standing. Consequently, people often respond defensively (i.e., derogating the source of unflattering bias feedback)—rather than increased awareness, egalitarian motivation, and prejudice regulation (e.g., Czopp et al., 2006; Howell et al., 2017)—when they receive feedback indicating bias (Howell et al., 2014; Vitriol & Moskowitz, 2021). These kinds of reactions can undermine antibias interventions (Dobbin & Kalev, 2016; Moskowitz & Vitriol, 2021). People who view themselves as having less explicit prejudice than others (Howell & Ratliff, 2017), hold incongruent implicit–explicit attitudes (Howell et al., 2014), lack awareness of bias (Perry et al., 2015), do not attend to their spontaneous affective reactions (Hahn & Gawronski, 2019), or who feel both morally impugned for having bias but perceive a lack of efficacy to control bias (Vitriol & Moskowitz, 2021), are especially defensive. Yet, it remains unknown how defensive reactions to bias feedback relate to the cognitive processes that underlie performance on measures of implicit attitudes.

Such an oversight is no small matter. Millions of visitors to the Project Implicit website receive feedback about the results of their performance on the IAT that seek to measure implicit attitudes toward many social categories. A major reason why such feedback is provided is to educate people about implicit social cognition and increase awareness of bias. While visitors to Project Implicit are often volunteers, real-world antibias interventions commonly encourage their participants to complete IATs on Project Implicit and elsewhere (e.g., Project Implicit, 2020). For many visitors to Project Implicit and participants in antibias interventions, the feedback is unflattering (e.g., when the direction and magnitude of their bias conflicts with egalitarian values). Defensive reactions to such feedback frustrate learning about bias (Moskowitz & Vitriol, 2021) and can undermine trust in psychological science and evidence-based strategies to reduce bias (see Dobbin & Kalev, 2016). Understanding the conditions under which defensiveness arises and its relationship to the cognitive processes that underpin performance on implicit measures can improve the effectiveness and accuracy with which educators communicate the science of social cognition and implicit bias. Indeed, such an investigation promises to help elucidate the psychological obstacles that undermine the kind of bias awareness that aids successful prejudice regulation and effective antibias interventions.

Quad Modeling and the Cognitive Underpinnings of Defensive Reactions

Implicit measures, like the IAT, were initially assumed to primarily reflect associations (e.g., in-group-good; Greenwald et al., 1998), but are now understood to also reflect the contributions of control-oriented processes that constrain the expression of biased associations (Calanchini & Sherman, 2013). Prior research examining defensive reactions to bias feedback has not revealed a consistent bivariate relationship between defensiveness and measures of implicit attitudes. However, this finding may reflect differing, and even countervailing, contributions of associations and control-oriented processes (e.g., Gonsalkorale et al., 2014). In the present research, we applied a method that disentangles the automatic and controlled processes that underpin responses on measures of implicit attitudes: the quadruple process model of implicit task performance (Quad model: Conrey et al., 2005). We then provide the first examinations of how the cognitive processes reflected in the Quad model independently predict defensive reactions to unflattering implicit racial bias feedback in a correlational analysis of two large samples (Sample 1, $N = 4{,}000$; Sample 2, $N = 4{,}000$) and an experimental investigation ($N = 547$) of the causal effect of bias feedback for the linkages between Quad parameters and defensiveness.

How might biased associations and control processes predict defensive responding to unflattering implicit racial bias feedback? One possibility is that defensiveness is related to biased associations. A positive relation might suggest that people are upset when their "hidden" biases are exposed. In contrast, a negative relation might suggest that people reject feedback that is discrepant from more flattering inferences about the content of their minds, or that people with more biased associations view bias feedback as an accurate characterization of their racial attitudes (Rivers & Hahn, 2019).

Another (not mutually exclusive) possibility is that defensiveness is related to control over biased associations. For example, people committed to egalitarian values are motivated and able to control stereotyping (Moskowitz et al., 1999). Failure to achieve a highly valued goal is threatening. A positive relation between defensiveness and control might indicate frustration that, despite exercising control over biased associations, people were unable to eradicate bias (hence the feedback). However, it is, in our view, more likely to be the case that low levels of control motivate defensive reactions to bias feedback. Vitriol and Moskowitz (2021) recently demonstrated that increasing perceived efficacy in controlling the expression of bias can reduce defensive reactions to bias feedback. More generally, Ruttan and Nordgren (2016) found that reduced self-regulatory capacity drives defensive information processing following negative feedback. Thus, defensiveness might be observed among people who struggle to exercise control over biased associations. In either case, the defensiveness–control relationship likely reflects the success or failure to regulate undesirable biased associations.

The Present Research

Across two studies, including a correlational analysis of two independent samples (Study 1) and an experimental manipulation of bias feedback (Study 2), we test the following predictions:

> *Hypothesis 1:* Higher levels of defensiveness will covary with lower levels of biased associations.
>
> *Hypothesis 2:* Higher levels of defensiveness will covary with lower levels of control over biased associations.
>
> *Hypothesis 3:* The relationship between Quad parameters and defensiveness described in H1 and H2 will only be observed among people who receive feedback about bias (vs. no feedback).

Hypotheses 1 and 2 are based on the results of the exploratory analysis in Study 1 and were preregistered in advance of the confirmatory analysis on a new sample (also Study 1). Neither the analysis nor predictions for Study 2 were preregistered. The preregistered procedure, hypotheses, and data analysis plan[1] for the confirmatory sample in Study 1 are available at https://osf.io/9yjqc/?view_only=6b5e72c9456447c1ad39ed25f494d04c (Vitriol et al., 2022).

Study 1

Transparency and Openness

We describe our sampling plan, all data exclusions (if any), all manipulations, and all measures in the study. We used RStudio to estimate Quad parameters and STATA Version 17.0 for statistical analyses. The hypotheses and analysis on the confirmatory sample in Study 1 were preregistered. Ethical approval was granted by the authors institutional ethics committee for all research activities reported in Study 1. All data and analysis code are available at https://osf.io/q264d/?view_only=ed84c0c937604b3591ef98f33077f500.

Furthermore, the 2013 Race IAT Codebook is available at https://osf.io/m8uxt.

Participants

From 358,896 visitors to the Project Implicit demonstration website (https://implicit.harvard.edu), we selected two independent samples of 4,000 White U.S. participants that fully completed the 2013 Black–White Implicit Association Test (Race IAT; Greenwald et al., 2003) made available by Project Implicit (Xu et al., 2014; see Table 1, for information about published papers that have utilized these data). The first sample was used for the exploratory analysis and the second sample was used for a preregistered confirmatory analysis, which was based on findings from the sample used in the exploratory analysis. Exploratory analysis, preregistration, and confirmatory analysis were all conducted in 2020, after the data were collected in 2013. Because both samples were drawn randomly from the full pool of participants, we had no knowledge of the responses among participants included in Study 1, prior to the random selection of our sample. Selections were further constrained to participants (a) who were 18 or over; (b) who responded to the explicit measures, including the demographic questions; (c) who did not have reaction times of less than 300 ms on more than 10% of IAT trials; (d) who had fewer than 30% of errors in the IAT; and (e) received the following feedback: "Your data suggest a slight/moderate/strong implicit preference for White People compared to Black People" (IAT scores of > .149).

Procedure, Materials, and Measures

Following informed consent, participants completed the Race IAT, then received personalized feedback based on their IAT performance, and subsequently answered a series of questions assessing their responses to the feedback (which we operationalize as defensiveness; described below). The IAT is a computer-administered categorization task (Greenwald et al., 1998; see https://osf.io/m8uxt, for the Race IAT materials). Participants are repeatedly presented with paired concepts and attributes, and faster response latencies indicate stronger implicit associations between the concepts. The Race IAT consists of two critical blocks; in one block, the labels "European Americans" and "Good" share the same response key, and "African Americans" and "Bad" share another response key (the congruent block for White participants). A trial involves a stimulus appearing at the center of the screen, which corresponds to one of the four labels (e.g., pictures of White or Black people; adjectives associated with "Good" or "Bad"), and the correct response key must be made before moving onto the next trial. In the other critical block, the instruction is reversed and the labels "European Americans" and "Bad" share the same response key and "African Americans" and "Good" share the same response key (incongruent block for White participants). If participants have faster reaction times in the congruent block relative to the incongruent block, this indicates a pro-White/anti-Black implicit attitude. The magnitude of this difference is reflected in a participant's *D*-scoring algorithm score (Greenwald et al., 2003; see Supplemental Materials, for further details).

Before completing the IAT, all participants received the following instructions:

> In this study you will complete an Implicit Association Test (IAT) in which you will be asked to sort pictures and words into groups as fast as you can. This study should take about 15 minutes to complete. At the end, you will receive your IAT result along with information about what it means.

After completing the Race IAT, participants were provided the following instructions:

[1] At the time of preregistration, the authors intended to analyze the defensiveness data dichotomously—an analytic choice reflecting the limitations of traditional process modeling methods. However, newer modeling methods allow us to analyze defensiveness continuously, which is what we report in this manuscript. Our methodology and analysis are otherwise consistent with all other aspects of the preregistered analysis plan for the confirmatory sample in Study 1. Similar results emerge for both sets of analyses.

Table 1

Data Transparency Table for All Papers That Utilized the Data From Study 1

Variable in the complete data set	Current article	MS 1	MS 2	MS 3	MS 4
Defensive responding	X				
Explicit racial attitudes			X	X	X
Race IAT *D* scores	X		X	X	X
Race IAT QUAD parameters	X	X	X		
Sexuality IAT QUAD parameters		X			
Age IAT QUAD parameters		X			
Gender-career IAT QUAD parameters		X			
Race-crime IAT QUAD parameters				X	
Disability IAT QUAD parameters		X			
Skin-tone IAT QUAD parameters		X			
Police use of lethal force				X	
Infectious disease metrics					X
Exposure to racial out-groups					X
Perceived vulnerability to disease					X

Note. MS = Manuscript; IAT = Implicit Association Test.

> Thank you for participating! The sorting test you just took is called the Implicit Association Test (IAT). You categorized pictures of Black and White people with Good and Bad words.

Based on their scores on this test, which were computed using the algorithm from Greenwald et al. (2003), all participants received one of seven possible forms of feedback about their performance on the IAT they completed, ranging from a strong (vs. moderate vs. slight) automatic preference for Black versus White people to a strong (vs. moderate vs. slight) automatic preference for White versus Black people with no preference for Black or White people as the mid-point:

> The results of this test indicate that you have [slight, moderate, strong] automatic preference in favor of White people compared to Black people

In the context of receiving feedback, participants are provided additional information about how to interpret the results of the IAT:

> These IAT results are provided for educational purposes only. The results may fluctuate and should not be used to make important decisions. The results are influenced by variables related to the test (e.g., the words or images used to represent categories) and the person (e.g., being tired, what you were thinking about before the IAT).

Participants are provided additional information about "how the IAT works," factors that can influence scores (e.g., order effects), the relationship between IAT scores and discrimination in applied contexts, a table containing the percent of web respondents who receive each score, and links to additional resources to learn more about implicit cognition and measurement.

After receiving feedback, participants completed a battery of measures (described below). All treatment of human participants in these experiments complied with American Psychological Association ethical standards. Table 2 reports the descriptive statistics and intercorrelations among assessed variables, for both the exploratory and confirmatory samples. All measures not included in the analyses are available in the study by Xu et al. (2014). A complete description of our measure of implicit attitudes is available in the Supplemental Materials. Here, we provide detailed information about our focal measure, defensiveness.

Defensiveness

We operationalize defensiveness here as derogation of the source of implicit bias feedback, which is similar to how researchers commonly conceptualized resistance to self-threatening information (Kunda, 1987; Sherman, 2013), counter-attitudinal messages (Tormala & Petty, 2004; Vitriol et al., 2019), and bias feedback (Vitriol & Moskowitz, 2021). Accordingly, we used the three-item index of defensiveness used in prior research examining defensive reactions to IAT feedback (i.e., Howell & Ratliff, 2017; Howell et al., 2014, 2017). The items for this measure included: "Whether I like my IAT score or not, it captures something important about me" (reverse coded), "The IAT reflects something about my automatic thoughts and feelings concerning this topic" (reverse coded), and "The IAT does not reflect anything about my thoughts or feelings, unconscious or otherwise" (exploratory: $\alpha = .80$, confirmatory: $\alpha = .79$). Participants responded to each item on a 4-point Likert scale: -2 = *strongly disagree*, -1 = *disagree*, 1 = *agree*, and 2 = *strongly agree*. We averaged across all three items, such that higher scores reflect more defensive responses. Responses to this measure were then rescaled to range from 0 to 1.

Results

The Quad Model

The Quad model estimates the independent contributions of four qualitatively distinct processes from responses on implicit measures using a multinomial processing tree model (see Conrey et al., 2005; Riefer & Batchelder, 1988). The Activation[2] parameter refers to the degree to which evaluative information (e.g., information connecting a stimulus with positive valence) is activated by the presence of a

[2] Previous Quad model research (e.g., Conrey et al., 2005) has referred to this parameter as the activation of associations. However, a growing body of research challenges the traditional assumption that implicit measures reflect information stored in simple associative networks (e.g., Amodio, 2019). Consequently, we use the label activation because it does not assume the representational structure underlying this cognitive process. Elsewhere in this manuscript, we use the term "associations" for linguistic convenience and in correspondence with existing social cognitive theory.

Table 2
Mean (SD) and Intercorrelations of Measures Assessed in Study 1

Variable	α	*M*	*SD*	IAT *D*	ACbb	ACwg	OB	*D*	*G*	Defensiveness
Exploratory sample										
IAT *D*	—	0.60	0.26	—						
ACbb	—	−1.70	0.05	0.39^{**}	—					
ACwg	—	−1.47	0.05	0.34^{**}	0.90^{**}	—				
OB	—	−0.25	0.27	-0.41^{**}	0.42^{**}	0.39^{**}	—			
D	—	1.37	0.07	0.14^{**}	-0.54^{**}	-0.61^{**}	-0.25^{**}	—		
G	—	0.16	0.03	$-0.03^{\dagger}$	-0.42^{**}	-0.28^{**}	-0.29^{**}	0.47^{**}	—	
Defensiveness	0.80	0.50	0.27	0.02	-0.05^{**}	-0.10^{**}	-0.15^{**}	0.02	0.22^{**}	—
Confirmatory sample										
IAT *D*	—	0.59	0.26	—						
ACbb	—	−1.71	0.29	0.15^{**}	—					
ACwg	—	−1.48	0.31	0.15^{**}	0.90^{**}	—				
OB	—	−0.18	0.80	-0.06^{**}	0.51^{**}	0.42^{**}	—			
D	—	1.37	0.40	−0.00	-0.58^{**}	-0.64^{**}	−0.00	—		
G	—	0.17	0.09	0.11^{**}	0.07^{**}	0.13^{**}	-0.18^{**}	0.22^{**}	—	
Defensiveness	0.79	0.50	0.26	−0.00	-0.15^{**}	-0.16^{**}	-0.24^{**}	0.02	-0.06^{**}	—

Note. Higher values indicate higher levels of the measured construct. The *G* parameter is the exception to this coding scheme: Higher values reflect a tendency to respond with "Good" attributes, and lower values reflect a tendency to respond with "Bad" attributes. Means (*SD*) of Quad values are probit-transformed. IAT *D* = participant's *D*-algorithm score on the Implicit Association Test; ACbb = activated Black-bad evaluations; ACwg = activated White-good evaluations; OB = overcoming bias; *D* = detection; *G* = guessing.
$^{*}p < .05$. $^{**}p < .01$. $^{\dagger}p < .10$.

stimulus on a given IAT trial. The more accessible the information (e.g., the stronger the association is between the in-group and "good" attributes), the more likely that information is to be activated and produce an evaluatively congruent response tendency. The detection parameter reflects the likelihood that the participant can discriminate between correct and incorrect responses, according to task requirements. Sometimes, activated evaluative information conflicts with the detected correct response. For example, on an IAT trial in which a picture of an outgroup member appears, and the outgroup and "good" stimuli share a response key (i.e., a normatively incongruent trial), the response tendency activated by the negatively evaluated outgroup (i.e., to press the button labeled "bad") conflicts with the required correct response (i.e., to press the button labeled "good"). In this case, according to the Quad model, overcoming bias navigates this conflict. The overcoming bias parameter refers to a control process that inhibits activated evaluative information from influencing behavior when this information conflicts with the required correct responses. Finally, the guessing parameter reflects any other processes that guide responses in the absence of influence from the other three parameters. The construct validity of the Quad model has been extensively demonstrated in previous research (see Calanchini & Sherman, 2013).

Parameter Estimation

To estimate the parameters specified in the Quad model, we employed the Bayesian approach proposed by Klauer (2010) to fit a multilevel extension of the model that treats participants and items as random factors for each model parameter (Judd et al., 2012), as implemented by the TreeBUGS *R* package (Heck et al., 2018). In this Bayesian approach, the T_1 statistic summarizes how well the model accounts for the pattern of observed response frequencies (i.e., correct and incorrect responses to Black, White, "Good," and "Bad" stimuli on compatible and incompatible trials) aggregated across participants within each condition (Klauer, 2010). This statistic corresponds to the goodness-of-fit chi-square statistic used in traditional modeling approaches (Riefer & Batchelder, 1988). The T_2 statistic summarizes how well the model accounts for the variances and correlations of these frequencies computed across participants, which quantifies how well the model accounts for individual differences between participants in the individual response frequencies (Klauer, 2010).

In the exploratory sample, participants made 7.14% errors. At the individual level, the median *p* value for T_1 was $p = .471$. At the group level, the observed versus predicted values for T_1 were .107 and .002, respectively, $p < .001$, and the observed versus predicted values for T_2 were 3.709 and .149, respectively, $p < .001$. In the confirmatory sample, participants made 7.11% errors. At the individual level, the median *p* value for T_1 was $p = .473$. At the group level, the observed versus predicted values for T_1 were .107 and .002, respectively, $p < .001$, and the observed versus predicted values for T_2 were 3.992 and .150, respectively, $p < .001$.

The nonsignificant *p* values for the individual-level statistics suggest that the Quad model provides good fit to these data, but the significant *p* values for the group-level statistics suggest that the observed outcomes differed significantly from the predicted outcomes. Relative to the group-level tests, the individual-level tests necessarily have less power to detect misfit. However, the group-level tests are based on an effective sample size of 4,000 participants × 120 IAT responses per participant = 480,000 observations each, so the group-level tests are highly powered to detect even a small degree of misfit. Thus, the significant *p* values indicate misfit but provide no insight into its degree. The T_1 statistic is chi-square-distributed, so we can calculate the effect size *w* to quantify the degree of misfit for each sample. For the exploratory sample $w = .008$, and for the confirmatory sample $w = .008$—both of which indicate a very small degree of misfit when controlling for sample size. The T_2 statistic is not chi-square-distributed, and there is no agreed-upon method to quantify degree of covariance fit in a way that controls for sample size. Consequently, we graphed the observed versus predicted frequencies and covariances, which we report in the Supplementary Materials. Visual inspection of these graphs indicates that differences between observed and predicted outcomes are minimal, which suggests that the Quad model provides a good fit for these data.

For each participant, we calculated two activation parameter estimates, and one estimate each for detection, overcoming bias, and guessing. One activation parameter reflected the extent to which positive White evaluations are activated, and the other activation parameter reflected the extent to which negative Black evaluations are activated. The guessing parameter was coded so that scores above .50 represented a bias toward responding with the "Good" key and scores below .50 represented a bias toward responding with the "Bad" key. Because parameters are estimated on the likelihood [0,1] scale, we probit-transformed them to approximate normal distributions for correlation and regression analyses.

Hypothesis Tests

Here we examined our primary hypotheses (see Tables 2–4). We do not observe a significant bivariate relationship between *D* scores from the IAT (IAT-*D*) and defensiveness in either sample (*p*s > .2). However, at the process level, we observe significant bivariate correlations between defensiveness and activation White-good (exploratory, $r = -.10$, $p < .001$; confirmatory, $r = -.16$, $p < .001$), activation Black-bad (exploratory, $r = -.04$, $p < .01$; confirmatory, $r = -.15$, $p < .001$), and overcoming bias (exploratory, $r = -.15$, $p < .001$; confirmatory, $r = -.24$, $p < .001$). Tables 3 and 4 summarize the results of a series of ordinary least squares regression, in which defensiveness was regressed on each Quad parameter, separately, for the exploratory and confirmatory analysis. Figure 1 graphically represents this pattern of results. Lower levels of activation White-good, activation Black-bad, and overcoming bias predicted increased levels of defensive responding to unflattering implicit bias feedback.

Study 2

In two large samples of White Americans, including a preregistered confirmatory replication of exploratory results, Study 1 demonstrates that higher levels of defensive responding covaries with lower levels of activation White-good, activation Black-bad, and overcoming bias (OB). Despite the strength of our evidence in Study 1, however, correlational data do not provide decisive evidence that Quad parameters caused defensive responding. While it is not logically possible for participants' defensive responding to feedback they receive after completing an IAT to retroactively influence their responses on the IAT, we cannot rule out the possibility that a third variable may account for both (a) higher levels of defensiveness and (b) lower levels of biased associations and control. To minimize this concern, in Study 2, we conducted an experiment in which participants were randomly assigned to receive no feedback or bias feedback, before completing a measure of defensiveness.

Transparency and Openness

Below we describe our sampling plan, all data exclusions (if any), all manipulations, and all measures in Study 2. Data were analyzed using STATA, Version 17.0. Ethical approval was granted by the authors institutional ethics committee for all research activities reported in Study 1. All data and analysis code are available at https://osf.io/q264d/?view_only=ed84c0c937604b3591ef98f33077f500.

Study Design

This study employed a single independent variable design (bias feedback prior to completing measures of defensiveness: yes, no). By manipulating feedback independent of performance on the IAT in an experimental study, we are able to examine the causal effect of feedback on the relationship between Quad parameters and defensive responding. Because we hypothesize that feedback causes defensiveness (see Howell et al., 2017; Vitriol & Moskowitz, 2021), we predict that the relationships between the Quad parameters and defensiveness observed in Study 1 will emerge only when participants receive feedback indicating bias (Hypothesis 3a). However, if people have some awareness of their biased associations while completing the IAT (Hahn et al., 2014), it is possible that we will observe a relationship between some of the Quad parameters and defensiveness, especially for the association parameters, without bias feedback, although to a lesser extent than when bias feedback is provided (Hypothesis 3b).

Participants

We used data from two independent samples that were recruited on Project Implicit (https://implicit.harvard.edu). In both samples,

Table 3
Defensiveness as a Function of Quad Parameters, Exploratory Sample

	Exploratory sample											
	Model 1				Model 2				Model 3			
	b	*SE*	95% CI	β	*b*	*SE*	95% CI	β	*b*	*SE*	95% CI	β
ACbb	−0.14**	0.05	−0.24, −0.04	−0.04**								
ACwg					−0.35***	.06	−0.46, −0.24	−0.10***				
OB									−.17***	.02	−0.21, −0.14	−0.15***
Intercept	−0.24**	.09	−0.40, −0.07		0.53***	.08	−0.69, −.36		−0.05***	.01	−0.08, −0.02	
F (dF)	7.69 (1, 3,998)				40.53 (1, 3,998)				89.84 (1, 3,998)			
R^2	0.002				0.010				0.022			

Note. Entries are ordinary least square unstandardized regression coefficients, with standard errors and confidence intervals. ACbb = activated Black-bad associations; ACwg = activated White-good associations; *SE* = standard error; CI = confidence interval; OB = overcoming bias.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. $^{\dagger}p < .10$.

Table 4

Defensiveness as a Function of Quad Parameters, Confirmatory Sample

	Confirmatory sample											
	Model 1				Model 2				Model 3			
	b	*SE*	95% CI	β	*b*	*SE*	95% CI	β	*b*	*SE*	95% CI	β
ACbb	−0.84***	.12	−1.07, −0.61	−.11***								
ACwg					−0.67***	.08	−0.83, −0.51	−.13***				
OB									−0.27***	.02	−0.30, −0.24	−.25 ***
Intercept	0.54***	.01	0.53, 0.56		0.55***	.01	0.54, 0.57		0.62***	.01	0.60, 0.63	
F (dF)	53.00 (1, 3,998)				68.96 (1, 3,998)				276.37 (1, 3,998)			
R^2	0.0131				0.0167				0.0647			

Note. Entries are ordinary least square unstandardized regression coefficients, with standard errors and confidence intervals. ACbb = activated Black-bad associations; ACwg = activated White-good associations; CI = confidence interval; *SE* = standard error; OB = overcoming bias.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. $^{\dagger}p < .10$.

data were initially collected from two IATs. We rely upon data only from participants who completed the Race IAT. Data collection for Study 2 preceded the preregistered confirmatory analysis in Study 1. However, data analysis for Study 2 occurred after the preregistered confirmatory analysis in Study 1.

Sample 2a

We recruited 395 U.S. citizens to complete a Race IAT for Study 2. Of these, 101 were dropped from analyses for having incomplete data ($n = 67$) or for identifying as non-White ($n = 34$). An additional 68 participants were excluded from analysis if their *D* scores on the Race IAT were below .15, consistent with the procedure used in Study 1. A total of 226 White U.S. participants were included in Sample 2a (73.5% females; mean age = 41.17, *SD* = 15.82). Most participants have earned at least a Bachelor's degree (48.9%).

Sample 2b

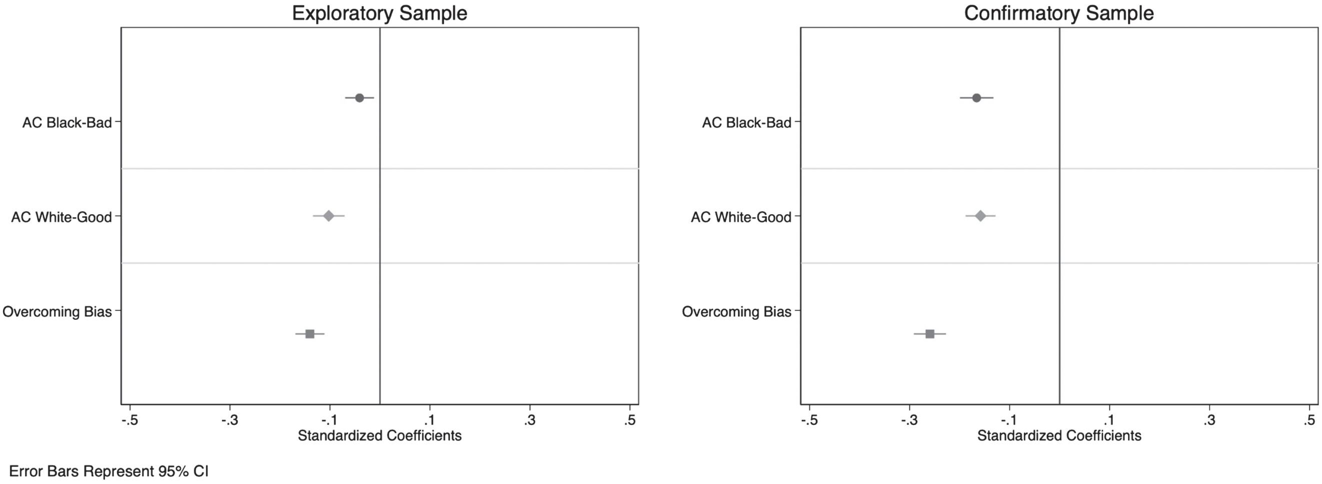
We recruited 482 White U.S. citizens to participate in Sample 2b. Of these, 100 were excluded from analyses for having incomplete data. An additional 103 participants were excluded from analysis if their *D* scores on the Race IAT were below .15, consistent with the procedure used in Study 1. A total of 321 White U.S. citizens completed the full survey (75.4% females; mean age = 35.56, *SD* = 15.45). Most participants have earned at least a Bachelor's degree (59.1%).

Statistical Power

With the combined sample size for Study 2 ($N = 547$), to detect an interaction between each Quad parameter and feedback condition, Study 2 had at least 65% power to detect a Cohen's *d* of .2 and 99% power to detect a Cohen's *d* of .5 or higher.

Figure 1

Effect of QUAD Parameters on Defensiveness



Note. Entries are ordinary least square standardized regression coefficients from Models 1, 2, and 3 in Tables 3 and 4. Each standard coefficient is from a single model estimating, separately, the relationship between defensiveness and each Quad parameter. Only AC (activation) and OB (overcoming bias) are shown here. Higher values indicate higher levels of the measured construct. CI = confidence interval.

Procedure

Participants were recruited from the Project Implicit educational website (https://implicit.harvard.edu/implicit/). All treatment of human participants in these experiments complied with American Psychological Association ethical standards. At the start of the study, participants reviewed a consent form and were then randomly assigned to one of two IAT conditions—Black–White IAT or insect–flower IAT. We use data only from participants who completed the Race IAT—the same IAT as used in Study 1. Feedback was provided in Study 2 in the same way as in Study 1. However, in Study 2, we experimentally manipulated whether participants completed our battery of measures before versus after receiving feedback. All participants were then debriefed and thanked for their time.

Study 2 reports a secondary analysis of data collected on two samples for a different study, which also involved the administration of multiple IATs. Both samples employed a 2 (Implicit Association Test; insect–flower vs. Black–White) × 2 (prefeedback evaluation; yes vs. no) between-subjects design. Participants were randomly assigned to complete one of two IAT and were then either assigned to complete measures of defensive responding before feedback or to do so after feedback. In Study 2, we used data from participants in the Race IAT condition.

Measures

Means (*SD*), alphas, and intercorrelations of all measures are shown in Table 5. The exact language used in the instructions, question stems, and items, for each sample in Study 2, are available in the Supplemental Materials.

Defensiveness

Participants in both samples responded to four items designed to measure defensiveness, but the language of the items and response scales varied slightly across the two samples (see the Supplemental Materials). Higher values indicate higher levels of defensiveness. As in Study 1, we rescaled responses to range from 0 to 1.

Results

Parameter Estimation

Parameters were estimated in the same way as in Study 1. However, we collapsed analyses across both samples to maximize statistical power, although the results are similar when analyzed separately. Participants made 5.18% errors. At the individual level, the median p value for T_1 was $p = .469$. At the group level, the observed versus predicted values for T_1 were .140 and .004, respectively, $p < .001$, and the observed versus predicted values for T_2 were 5.485 and .508, respectively, $p < .001$. The magnitude of misfit in terms of T_1 was very small, $w = .009$, and visual inspection of the model fit graphs (see Supplemental Materials) indicates that differences between observed and predicted outcomes are minimal, which suggests that the Quad model provides good fit to these data.

Hypothesis Tests

First, we examine the bivariate relationships among IAT *D* scores, activation White-good, activation Black-bad, or overcoming bias and defensiveness, separately, for Samples 2a and 2b, collapsed across feedback condition. As reported in Table 5, and consistent with H3, when collapsing across feedback conditions, we do not observe significant bivariate correlations between defensiveness and IAT *D* scores (Sample 1a, $r = -.05$, $p > .10$; Sample 1b, $r = -.01$, $p > .10$), activation White-good (Sample 1a, $r = .02$, $p > .10$; Sample 1b, $r = .00$, $p > .10$), activation Black-bad (Sample 1a, $r = .01$, $p > .10$, Sample 1b, $r = .02$, $p > .10$), and overcoming bias (Sample 1a, $r = -.03$, $p > .10$; Sample 1b, $r = -.04$, $p > .10$).

Second, we examine the extent to which the relationship between IAT *D* scores, activation White-good, activation Black-bad, or

Table 5
Mean (SD) and Intercorrelations of Measures Assessed in Study 2

Variable	α	*M*	*SD*	IAT *D*	ACbb	ACwg	OB	*D*	*G*	Defensiveness
Sample 2a										
IAT *D*	—	0.57	0.24	—						
ACbb	—	−1.89	0.40	0.16*	—					
ACwg	—	−1.68	0.33	0.08	0.95**	—				
OB	—	0.27	0.65	−0.02	0.49**	0.47**	—			
D	—	1.75	0.41	0.06	−0.45**	−0.49**	−0.13*	—		
G	—	0.16	0.07	0.04	−0.32**	−0.32**	−0.46**	0.34**	—	
Defensiveness	0.86	0.36	0.24	−0.05	0.02	0.02	−0.03	−0.02	0.06	—
Sample 2b										
IAT *D*	—	0.60	0.27	—						
ACbb	—	−1.88	0.41	0.17**	—					
ACwg	—	−1.67	0.34	0.18**	0.94**	—				
OB	—	0.31	0.61	−0.05	0.41**	0.40**	—			
D	—	1.70	0.47	−0.05	−0.50**	−0.53**	−0.06	—		
G	—	0.16	0.08	−0.06	−0.35**	−0.33**	−0.44**	0.26**	—	
Defensiveness	0.89	0.37	0.22	−0.01	0.02	0.00	−0.04	−0.10†	0.01	—

Note. IAT *D* = participant's *D*-algorithm score on the Implicit Association Test; ACbb = activated Black-bad evaluations; ACwg = activated White-good evaluations; OB = Overcoming Bias; *D* = Detection; *G* = Guessing. Higher values indicate higher levels of the measured construct. The *G* parameter is the exception to this coding scheme: Higher values reflect a tendency to respond with "Good" attributes, and lower values reflect a tendency to respond with "Bad" attributes. Means(*SD*) of Quad values are probit-transformed.
* $p < .05$. ** $p < .01$. † $p < .10$.

overcoming bias and defensiveness is moderated by feedback conditions. To do so, we tested for an interaction between feedback condition (using a dummy coded variable in which 0 = no feedback, 1 = feedback), activation White-good, activation Black-bad, or overcoming bias, separately. For all analyses described below, we estimate a multilevel model with maximum likelihood estimation and with sample submitted as a random-intercept term.

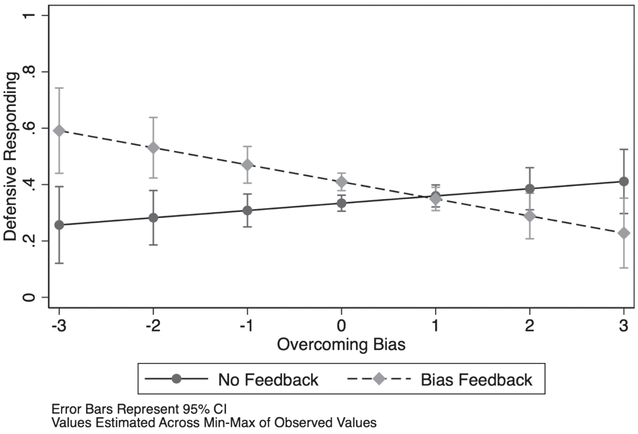
The interaction obtained significance for overcoming bias, $b = -.09$, CI 95% [−.15, −.03], $p = .005$, but not for activation White-good, $b = -.11$, CI 95% [−.22, .00], $p = .057$, or activation Black-bad, $b = -.08$, CI 95% [−.17, .02], $p = .106$. To decompose the significant interaction for overcoming bias, we first regressed defensive responding on overcoming bias for participants in the bias feedback or no feedback condition, separately. We observe a significant relationship between defensive responding and overcoming bias in the bias feedback, $b = -0.06$, CI 95% [−.11, −.01], $p = .011$, but not in the no feedback, $b = .03$, CI 95% [−.01, .07], $p = .20$, conditions. We are also interested in whether the effect of the feedback is differentially impactful at high and low levels of Overcoming Bias. To examine this relationship, we tested the effect of bias feedback at 1 *SD* above and below the mean of overcoming bias pooled across both samples. The effect of bias (vs. no) feedback obtained significance at 1 *SD* below, $b = .10$, CI 95% [.05, .15], $p < .001$, and above, $b = .05$, CI 95% [.01, .09], $p < .01$, the mean of OB. Inspection of the coefficients for both sets of simple slope analyses indicates that the size of the effect of bias feedback on defensiveness was significant across all levels of overcoming bias, but, importantly, was larger at lower (vs. higher) levels of overcoming bias. Thus, we obtain evidence in support of Hypothesis 3: bias feedback (vs. no feedback) caused increased levels of defensiveness among participants low (vs. high) in overcoming bias. Figure 2 graphically represents the relationship between the overcoming bias parameter and defensiveness, separately for participants in the bias feedback and no feedback conditions.

Discussion

We report the results of one of the first investigations of if or how the cognitive processes that underpin performance on implicit measures predict defensive reactions to unflattering implicit racial bias feedback. By using the Quad model, the present research helps to answer an unaddressed question of theoretical and practical importance: How are biased associations and control processes related to defensive responding among people who receive unflattering implicit racial bias feedback?

Figure 2

Effect of Feedback × Overcoming Bias on Defensiveness



Note. CI = confidence interval.

We find that similarities in IAT *D* scores across levels of defensiveness can obscure differences in underlying cognitive processes. Study 1 relied on correlational analyses across exploratory and confirmatory samples, and demonstrated that people who responded more (vs. less) defensively had less control over their biased associations. Study 2 provides particularly strong evidence for a link between control and defensive responding, as the relationship was limited to conditions under which participants received bias (vs. no) feedback. One interpretation of these results is that people are defensive when they try but learn they have failed to control their biases. Control can be exerted through implicit processes in which bias awareness (via negative feedback) triggers an egalitarian goal outside of conscious awareness. These goals can implicitly inhibit incompatible constructs, such as stereotypes, before they ever become accessible (Moskowitz & Li, 2011; Moskowitz et al., 2000; Moskowitz & Stone, 2012). Control can also be exerted through deliberate attempts to focus on egalitarian goals and regulate the expression of bias following increased awareness of having acted with bias or prejudice that conflicts with importantly held egalitarian values or goals (Monteith et al., 2001). Regardless of which route to control people might employ, feedback that highlights their failure to achieve egalitarian goals does not always motivate increased commitment to egalitarian goals; often, the inability to achieve one's egalitarian goals can instead motivate self-protective orientations (Moskowitz & Vitriol, 2021; Shepperd et al., 2008), which may be particularly true among people who value fairness and aspire to be nonprejudiced (Howell et al., 2014). Our findings align with this interpretation and may help explain why intervention programs that aim to create awareness of bias are not always successful in reducing bias (Lai et al., 2016). That is, people may feel unable to recognize or control bias, and, consequently, become unwilling to learn about the existence of bias and what skills, strategies, or policies they can implement to reduce it.

Critically, the finding that bias feedback leads to defensiveness among people who show low levels of control points to a promising approach for raising awareness in a way that minimizes resentment, anger, denial, and backlash (e.g., Howell & Ratliff, 2017): to increase perceived efficacy in control over the expression and consequence of bias. A recent investigation examining strategies for reducing defensive reactions to unflattering bias feedback supports this prediction. Vitriol and Moskowitz (2021) demonstrate that increasing perceived efficacy in recognizing and controlling bias is an effective strategy for reducing defensive reactions to bias feedback. By reducing defensiveness, this intervention not only promoted awareness of bias that persisted 6 months after the original feedback was provided, but also increase support for antiprejudice interventions and policies, reduced endorsement of negative racial stereotypes, and motivated increased commitment to egalitarian goals. The results of the present research further reinforce the value of adopting such an approach in antibias interventions to help mitigate defensiveness, increase openness to the science of implicit cognition, promote awareness of the mechanics of one's own mind, and motivate egalitarian behavior.

We observe an inconsistent pattern of results for the relationship between biased associations and defensiveness. In the correlational analyses of Study 1, people who responded more (vs. less) defensively also had weaker, but nonzero, White-good and Black-bad associations. This finding joins recent research challenging traditional assumptions about the unconscious nature of the IAT. People can sense their performance on implicit measures (Monteith et al., 2001) and can sometimes even predict its outcome (Hahn et al., 2014), especially when they reflect on their biased associations (Rivers & Hahn, 2019). From this perspective, the results of Study 1 suggest that defensiveness may arise, in part, from inconsistency between the feedback and introspective inferences about biased associations. Indeed, Howell and Ratliff (2017) find that defensive reactions to unflattering bias feedback are particularly common among people who explicitly hold egalitarian attitudes. This finding also raises the alternative possibility that the negative relationship between biased associations and defensiveness may be driven by increased acceptance of the feedback among participants with high levels of biased associations. Perhaps such participants are not surprised nor threatened to learn of their own bias and may even regard the feedback as consistent with their own self-concepts (e.g., Moskowitz & Vitriol, 2021; Talaifar & Swann, 2017).

Future research can more directly test these explanations by investigating the extent to which the link between defensiveness and White-good and Black-bad associations are conditioned on egalitarian motivations, explicit attitudes, and the discrepancy between implicit and explicit attitudes. For example, people who value egalitarianism may be both (a) less likely to hold biased associations and (b) more likely to be threatened by feedback indicating that they have failed to act in ways consistent with their egalitarian beliefs. As our findings demonstrate, a *D* score indicating bias can still arise among respondents who fail to control (even low levels of) biased associations. Thus, participants with low levels of explicit bias may have low levels of biased associations, but because they failed to exercise control, they nonetheless produced biased *D* scores, and receive bias feedback, toward which they are defensive. Additionally, the negative relationship between biased associations and defensiveness may only characterize people with high levels of explicit bias, who regard the feedback to be accurate and consistent with their worldview and hence are not defensive toward bias feedback. Examining the moderating role of explicit bias and egalitarian motivations is a natural and valuable extension of this work, as it would help explain the extent to which these possibilities can account for our pattern of results.

Despite the consistent pattern of correlations for biased associations in Study 1, we did not observe a significant interaction between bias feedback (vs. no feedback) and White-good or Black-bad associations on defensive responding in Study 2. One possible explanation for these discrepant results is that we lacked the statistical power to detect the effect in Study 2. Indeed, the magnitude of the overcoming bias–defensiveness relationship is substantially larger than that for activation White-good and activation Black-bad, and we had only 65% statistical power to detect small effects. Additional research with larger samples is needed to clarify the causal effect of bias feedback on this dynamic.

Despite the strength of our evidence, the characteristics of visitors to Project Implicit are not representative of the general population. However, millions of people visit Project Implicit annually. Thus, the results of the current research have practical real-world implications even if our observations do not generalize beyond our samples. Future research would nonetheless benefit from more direct investigation of external validity (cf. Vitriol et al., 2019). Doing so will advance our understanding of the psychological obstacles that undermine the kind of bias awareness that aids successful prejudice regulation and effective antibias interventions.

Indeed, it is a scientific imperative to understand what contributes to and can reduce defensive reactions to the existence of inequality and discrimination, and its linkages to psychological sources of bias, representing a critical challenge to our field and to those who desire a more equitable and just society. We write this article at a time in which backlash against the teaching and learning about the concept of implicit bias has gained mainstream political traction and momentum. For example, in October 2020, former President Trump signed an executive order prohibiting federal agencies from implementing training intended to combat bias and inequality in the legal system, describing it as "anti-American" (Guyn, 2020; Office of Management & Budget, 2020). The National Association for the Advancement of Colored People Legal Defense Fund described this effort as an "extraordinary and unprecedented act … to undermine efforts to foster diversity and inclusion in the workplace" (National Association for the Advancement of Colored People, 2020). Several U.S. States—including Florida, Idaho, Louisiana, North Carolina, Oklahoma, Rhode Island, Tennessee, and Texas—have recently banned the teaching of "divisive concepts" about racial inequality and diversity, such as implicit bias (Adams, 2021; Cineas, 2021; Florido, 2021; Samee Ali, 2021). Thus, revealing the psychological underpinning of defensiveness toward implicit bias is not only consequential for the development of effective forms of antibias education, but also may be critical to the realization of a more fair, equitable, and just society. We hope that our investigation represents a modest step in this more egalitarian direction.

References

Adams, C. (2021). How Trump ignited the fight over critical race theory in schools. *NBCNews*. https://www.nbcnews.com/news/nbcblk/how-trump-ignited-fight-over-critical-race-theory-schools-n1266701

Amodio, D. M. (2019). Social Cognition 2.0: An interactive memory systems account. *Trend in Cognitive Sciences*, *23*(1), 21–33. https://doi.org/10.1016/j.tics.2018.10.002

Axt, J. R., Casola, G., & Nosek, B. A. (2019). Reducing social judgment biases may require identifying the potential source of bias. *Personality and Social Psychology Bulletin*, *45*(8), 1232–1251. https://doi.org/10.1177/0146167218814003

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. https://doi.org/10.1016/j.jesp.2017.06.003

Calanchini, J., & Sherman, J. (2013). Implicit attitudes reflect associative, non-associative, and non-attitudinal processes. *Social and Personality Psychology Compass*, *7*(9), 654–667. https://doi.org/10.1111/spc3.12053

Carter, E. C., Onyeador, I. N., & Lewis, N. A., Jr. (2020). What do we know about (implicit) bias and what does it mean for bias reduction training? *Behavioral Science & Policy*, *6*(1), 57–70. https://doi.org/10.1353/bsp.2020.0005

Cineas, F. (2021). Critical race theory, and trump's war on it, explained. *Vox*. https://www.vox.com/2020/9/24/21451220/critical-race-theory-diversity-training-trump

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition:

The quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*(4), 469–487. https://doi.org/10.1037/0022-3514.89.4.469

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology*, *90*(5), 784–803. https://doi.org/10.1037/0022-3514.90.5.784

Dobbin, F., & Kalev, A. (2016). Why diversity programs fail. *Harvard Business Review*, *94*(7). https://stratserv.co/wp-content/uploads/2020/08/Why-Diversity-Programs-Fail.pdf

Florido, A. (2021). Teachers say laws banning Critical Race Theory are putting a chill on their lessons. *National Public Radio*. https://www.npr.org/2021/05/28/1000537206/teachers-laws-banning-critical-race-theory-are-leading-to-self-censorship

Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2014). Measures of implicit attitudes may conceal differences in implicit associations: The case of antiaging bias. *Social Psychological & Personality Science*, *5*(3), 271–278. https://doi.org/10.1177/1948550613499239

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

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. https://doi.org/10.1037/0022-3514.85.2.197

Guyn, J. (2020). Donald Trump executive order banning diversity training blocked by federal judge. *USA Today*. https://www.usatoday.com/story/money/2020/12/23/trump-diversity-training-ban-executive-order-blocked-federal-judge/4033590001/

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*(5), 769–794. https://doi.org/10.1037/pspi0000155

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369–1392. https://doi.org/10.1037/a0035028

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*(1), 264–284. https://doi.org/10.3758/s13428-017-0869-7

Hillard, A. L., Ryan, C. S., & Gervais, S. J. (2013). Reactions to the implicit association test as an educational tool: A mixed methods study. *Social Psychology of Education*, *16*(3), 495–516. https://doi.org/10.1007/s11218-013-9219-5

Howell, J. L., Gaither, S. E., & Ratliff, K. A. (2014). Caught in the middle: Defensive responses to IAT feedback among whites, blacks, and biracial black/whites. *Social Psychological & Personality Science*, *6*(4), 373–381. https://doi.org/10.1177/1948550614561127

Howell, J. L., & Ratliff, K. A. (2017). Not your average bigot: The better-than-average effect and defensive responding to implicit association test feedback. *British Journal of Social Psychology*, *56*(1), 125–145. https://doi.org/10.1111/bjso.12168

Howell, J. L., Redford, L., Pogge, G., & Ratliff, K. A. (2017). Responding defensively to IAT feedback. *Social Cognition*, *35*(5), 520–562. https://doi.org/10.1521/soco.2017.35.5.520

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, *53*(4), 636–647. https://doi.org/10.1037/0022-3514.53.4.636

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. https://doi.org/10.1037/xge0000179

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. https://doi.org/10.1037/0022-3514.65.3.469

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, *19*(4), 395–417. https://doi.org/10.1521/soco.19.4.395.20759

Moskowitz, G., & Vitriol, J. A. (2021). A social cognition model of bias reduction. In A. Nordstrom & W. Goodfriend (Eds.), *Innovative stigma and discrimination reduction programs*. Routledge. https://doi.org/10.4324/9781003042464-1

Moskowitz, G. B. (2010). On the control over stereotype activation and stereotype inhibition. *Social and Personality Psychology Compass*, *4*(2), 140–158. https://doi.org/10.1111/j.1751-9004.2009.00251.x

Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, *77*(1), 167–184. https://doi.org/10.1037/0022-3514.77.1.167

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. https://doi.org/10.1016/j.jesp.2010.08.014

Moskowitz, G. B., Salomon, A. R., & Taylor, C. M. (2000). Preconsciously controlling stereotyping: Implicitly activated egalitarian goals prevent the activation of stereotypes. *Social Cognition*, *18*(2), 151–177. https://doi.org/10.1521/soco.2000.18.2.151

Moskowitz, G. B., & Stone, J. (2012). The proactive control of stereotype activation: Implicit goals to not stereotype. *Zeitschrift für Psychologie*, *220*(3), 172–179. https://doi.org/10.1027/2151-2604/a000110

National Association for the Advancement of Colored People. (2020). *NAACP legal defense fund, National Urban League, National Fair Housing Alliance file suit against Trump administration; African American Policy Forum launches #TruthBeTold campaign*. NAACP Legal Defense and Educational Fund. https://www.naacpldf.org/wp-content/uploads/EO-Complaint__Press-Release-10.29.20-FINAL.pdf

Office of Management and Budget. (2020). *Training in the federal government*. https://www.whitehouse.gov/wp-content/uploads/2020/09/M-20-34.pdf

Parker, L. R., Monteith, M. J., Moss-Racusin, C. A., & Van Camp, A. R. (2018). Promoting concern about gender bias with evidence-based confrontation. *Journal of Experimental Social Psychology*, *74*, 8–23. https://doi.org/10.1016/j.jesp.2017.07.009

Perry, S. P., Murphy, M. C., & Dovidio, J. F. (2015). Modern prejudice: Subtle, but unconscious? The role of bias awareness in Whites' perceptions of personal and others' biases. *Journal of Experimental Social Psychology*, *61*, 64–78. https://doi.org/10.1016/j.jesp.2015.06.007

Project Implicit. (2020, March). *Products and services*. https://www.projectimplicit.net/services.html

Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T., & Huguet, P. (2019). Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour*, *3*(11), 1171–1179. https://doi.org/10.1038/s41562-019-0686-3

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339. https://doi.org/10.1037/0033-295X.95.3.318

Rivers, A. M., & Hahn, A. (2019). What cognitive mechanisms do people reflect on when they predict IAT scores? *Personality and Social Psychology Bulletin*, *45*(6), 878–892. https://doi.org/10.1177/0146167218799307

Ruttan, R. L., & Nordgren, L. F. (2016). The strength to face the facts: Self-regulation defends against defensive information processing. *Organizational Behavior and Human Decision Processes*, *137*, 86–98. https://doi.org/10.1016/j.obhdp.2016.06.006

Samee Ali, S. (2021). Florida board of education passes rule banning critical race theory in classrooms. *NBC News*. https://www.nbcnews.com/news/us-news/florida-board-education-passes-rule-banning-critical-race-theory-classrooms-n1270363

Shepperd, J., Malone, W., & Sweeny, K. (2008). Exploring causes of the self-serving bias. *Social and Personality Psychology Compass*, *2*(2), 895–908. https://doi.org/10.1111/j.1751-9004.2008.00078.x

Sherman, D. K. (2013). Self-affirmation: Understanding the effects. *Social and Personality Psychology Compass*, *7*(11), 834–845. https://doi.org/10.1111/spc3.12072

Talaifar, S., & Swann, W. B., Jr. (2017). Self-verification theory. In V. Zeigler-Hill & T. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp 1–9). Springer. https://doi.org/10.1007/978-3-319-28099-8_1180-1

Tormala, Z. L., & Petty, R. E. (2004). Source credibility and attitude certainty: A metacognitive analysis of resistance to persuasion. *Journal of Consumer Psychology*, *14*(4), 427–442. https://doi.org/10.1207/s15327663jcp1404_11

Vitriol, J. A., Larsen, E. G., & Ludeke, S. G. (2019). The generalizability of personality effects in politics. *European Journal of Personality*, *33*(6), 629–701. https://doi.org/10.1002/per.2222

Vitriol, J. A., & Moskowitz, G. (2021). Reducing defensive responding to implicit bias feedback: On the role of perceived moral threat and efficacy to change. *Journal of Experimental Social Psychology*, *96*, Article 104165. https://doi.org/10.1016/j.jesp.2021.104165

Vitriol, J. A., O'Shea, B., & Calanchini, J. (2022, August 8). *Defensive reactions and QUAD*. https://osf.io/q264d

Xu, K., Nosek, B., & Greenwald, A. G. (2014). Psychology data from the race implicit association test on the project implicit demo website. *Journal of Open Psychology Data*, *2*(1). Article e3. https://doi.org/10.5334/jopd.ac

Received August 30, 2022
Revision received November 18, 2022
Accepted January 14, 2023 ■