Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 11, 2859–2878
https://doi.org/10.1037/xge0001596

Eliciting Cognitive Consistency Increases Acceptance of Implicit Bias

Joseph A. Vitriol and Mahzarin R. Banaji
Department of Psychology, Harvard University

Resistance to knowledge about implicit bias jeopardizes the ability to learn, understand, and act to outsmart bias. Across three experiments and five independent samples ($N > 3{,}500$), conditions that increase cognitive consistency were created alongside control conditions. In Experiment 1, using a race (Black–White) Implicit Association Test (IAT), cognitive consistency was enhanced when participants evaluated the validity and utility of the test before, rather than after, receiving the test result, leading to greater acceptance of bias. In Experiments 2 and 3, participants either evaluated their performance on a Black–White IAT alone or evaluated their performance on a morally innocuous Insect–Flower IAT prior to a Black–White IAT. Again, resistance to evidence of implicit racial bias was reduced in the latter condition, where the imperative for cognitive consistency was heightened. In all three experiments, creating ordinary conditions to heighten cognitive consistency was associated with increased bias awareness and acceptance and, additionally, with support for actions to minimize its consequence–outcomes critical to achieving effective bias education.

Public Significance Statement

To be objective and to demonstrate cognitive consistency between attitudes, beliefs, or behaviors, is lauded. However, human behavior, as psychologists have discovered, often tells a different story. In these studies, we show that implementing simple techniques, such as evaluating a test prior to learning one's test result or completing a morally innocuous test prior to a morally significant one, facilitates greater awareness and acceptance of implicit bias, and an increased willingness to learn about sources of error in judgment and decisions. Because these results derive from basic interventions that are easy to implement, these results (besides identifying psychological mechanisms that produce awareness of bias) may also be useful in developing effective education about the science of implicit bias.

Keywords: bias awareness and education, cognitive consistency, implicit bias, Implicit Association Test, race bias

Supplemental materials: https://doi.org/10.1037/xge0001596.supp

In the late 1930s, when Gunnar Myrdal, a Swedish sociologist, interviewed southern Americans about their views on race, he was struck by their acknowledgment of, and even distress about, the discrepancy between their commitment to values of freedom and equality on the one hand, and the history and legacy of slavery in the United States on the other hand (Myrdal, 1944). This observation was so pivotal to Myrdal's analysis that he titled his two-volume set, An American Dilemma. Fifty years later, by using established methods to study automatic or implicit cognition, psychologists (e.g., Devine, 1989) discovered what today we might label the New American Dilemma: a curious disparity between an individual's own consciously held attitudes and beliefs (e.g., an avowed belief in racial equality) and their less conscious, automatic or implicit attitudes and beliefs. Today, a large body of research provides evidence of thoughts and feelings that operate without conscious awareness or conscious control, and that deviate from the individual's own espoused intentions and values (Banaji & Greenwald, 2013; Gawronski, 2019; see Morehouse & Banaji, 2024 for a

This article was published Online First May 23, 2024.
Sarah Gaither served as action editor.
Joseph A. Vitriol https://orcid.org/0000-0002-2715-6123
Joseph A. Vitriol is now at Department of Management, Lehigh University College of Business.
The data and ideas of this article have been presented at the 83rd Academy of Management Meeting in Boston, Massachusetts, United States, in August 2023; the Annual Meeting of the Society for Personality and Social Psychology in New Orleans, Louisiana, United States, in February 2020; the 31st Annual Convention of the Association for Psychological Science in Washington, District of Columbia, United States, in 2019; the Behavioral Sciences and Leadership Department at the United States Air Force Academy; the Management Department of the Business College at Lehigh University; the Psychology Department at Occidental College; the Behavioral Science and Law Department of the New College at Arizona State University; the Social Cognition Lab in the Psychology Department at the University of California, Riverside, California, United States; and the Implicit Social Cognition Lab in the Psychology Department at Harvard University. Data and data syntax are available at https://osf.io/7tfcs/?view_only=0819a809d2964aec8fe44ff998db601c.

Joseph A. Vitriol served as lead for data curation, formal analysis, and writing–original draft and served in a supporting role for methodology. Mahzarin R. Banaji served as lead for methodology and contributed equally to writing–original draft. Joseph A. Vitriol and Mahzarin R. Banaji contributed equally to conceptualization and writing–review and editing.

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

comprehensive review of implicit race bias based on 15 years of Implicit Association Test [IAT] data).

Major theories in psychology hold that awareness is necessary, albeit insufficient, for the successful regulation of implicit bias (e.g., Bezrukova et al., 2016; Czopp et al., 2006; Lai et al., 2016; Moskowitz & Li, 2011; Perry et al., 2015). Education is important to achieving this goal, as we showed in a recent case involving the education of a police department whose members were not entirely welcoming of such education at the start (Vitriol, Banaji, & Lowe, 2024). By creating awareness of implicit bias, education can create the conditions necessary for change. This is a reasonable expectation given that many people already believe in fairness and equality (Bergsieker et al., 2010; Moskowitz, 2002), and also desire to act in a manner that is consistent with their values. Indeed, research has pointed out that awareness of inconsistency between egalitarian values and behaviors can increase commitments to fairness and attempts to mitigate bias in one's actions (e.g., Aneja et al., 2023). Thus, sustainable reduction of bias in real-world contexts often, but not always, follows increased awareness of bias (Adams et al., 2014; Axt et al., 2019; Bezrukova et al., 2016; Burns et al., 2017; Carnes et al., 2015; Devine et al., 2012; Forscher et al., 2017; Monteith, 1993; Monteith & Mark, 2009; Moskowitz et al., 1999; Parker et al., 2018; Regner et al., 2019).

However, efforts to educate about implicit bias encounter a unique obstacle (Moskowitz & Vitriol, 2021). The act of learning about bias can itself elicit resistance to the evidence. If one believes oneself to be a moral actor behaving in accordance with consciously endorsed values (O'Brien et al., 2010; Sommers & Norton, 2006), it is challenging to accept evidence to the contrary. As such, resistance is common when evidence of implicit bias is presented (Howell et al., 2015; Vitriol & Moskowitz, 2021). This resistance, in turn, can undermine learning and growth, as has been observed when people perceive that they have acted inconsistently with their own stated beliefs and values (Aronson, 1999; Howell et al., 2013; Stone, 2001).

The current program of work tests a novel strategy for increasing awareness of bias. We leveraged the motivation for cognitive consistency to test whether it would lead to acceptance of bias. In three experiments, we created conditions that elicited a favorable view of a test of implicit attitudes before providing feedback about bias. We tested whether this acknowledgment subsequently increased awareness of bias, acceptance of bias, and commitment to egalitarian goals.

Reactions to Bias Education

There are many paths to discounting evidence of one's bias, and a common one is to challenge the validity of the instrument that provides unflattering feedback. Evidence supporting the idea that humans look to maintain a positive view of the self is as old as the field of social psychology itself (Banaji et al., 1993; Greenwald, 1980; O'Brien et al., 2010). Providing evidence of bias can ironically make people less open to mitigating or regulating bias (Duguid & Thomas-Hunt, 2015; Legault et al., 2011; Moskowitz & Vitriol, 2021). Feedback about or evidence of bias can challenge egalitarian self-concepts (Dovidio & Gaertner, 2000; Frantz et al., 2004), undermine belief in the objectivity of one's judgments (e.g., "bias blindspots"; Pronin, 2007), and threaten the perceived fairness of the status quo (Jost et al., 2004).

A literature on the limitations and challenges of raising awareness of bias is burgeoning (Howell & Ratliff, 2016; Howell et al., 2017; Moskowitz & Vitriol, 2021; Onyeador et al., 2021; Rothman et al., 2022; Vitriol, O'Shea, & Calanchini, 2024). This work demonstrates that defensive responding (i.e., rejection of the information and indeed the underlying science as "fake") in response to evidence of bias is relatively common (Czopp et al., 2006; Hillard et al., 2013; Howell et al., 2015). Unlike the experience of self-directed negative affect or guilt (e.g., Monteith, 1993; Monteith et al., 1993, 2002), resistance reduces support for antibias interventions and undermines motivations for prejudice regulation. In this way, awareness-raising strategies for bias reduction may backfire (e.g., Dobbin & Kalev, 2016; Duguid & Thomas-Hunt, 2015; Legault et al., 2011; Perry et al., 2015). While of significant importance for effective bias education, however, efforts to identify strategies to increase acceptance of implicit bias feedback have largely been understudied (see Rothman et al., 2022; Vitriol & Moskowitz, 2021, for exceptions).

Leveraging Consistency Motivations for Bias Education

Once people commit to a judgment or perform a behavior, it constrains subsequent thoughts and actions over time (Gawronski, 2012). For example, people may revise their attitudes to align with freely chosen past actions (e.g., Festinger & Carlsmith, 1959). Stone et al. (1994) provide a particularly compelling demonstration of the power of commitment in their work on hypocrisy induction, which makes salient a conflict between current belief and past action; a conflict commonly resolved through behavioral change in the direction of the former. Similarly, Cialdini (2008) describes a wide range of strategies for social influence that leverage initial commitment to increase subsequent persuasion and compliance (i.e., foot-in-the-door and low-ball technique). More generally, Jost et al. (2004) argue that people are motivated to defend and justify the status quo to maintain a preexisting belief in the fairness and desirability of existing social structures and power dynamics. Even when people lack a meaningful basis for discerning their internal states or prior beliefs, they may infer their attitudes based on observations of their past behavior (e.g., Bem, 1972).

What these classic findings and perspectives suggest is that by asking people to evaluate the validity of a measure of implicit attitudes before receiving unflattering feedback, they may be less resistant and more willing to regard that feedback, once provided, as credible. These prefeedback commitments are expected to put into motion psychological processes of cognitive consistency that yield awareness and even acceptance of bias. We investigate this possibility on the Project Implicit website (https://implicit.harvard.edu). A diversity of visitors to the site regularly complete IATs (Greenwald et al., 2009) on topics of their choosing and then receive feedback about their implicit attitudes. Since its inception, more than 30 million tests have been completed on the Project Implicit site. This venue provided a suitable context to explore the power of cognitive consistency to shape awareness and acceptance of bias.

The Current Research

Across five independent samples and three experimental paradigms ($N > 3{,}500$), we tested the effectiveness of a novel strategy for increasing bias awareness by leveraging the tendency for cognitive consistency to reduce resistance to objective evidence of bias.

In all experiments, we sampled White U.S. citizens who visited Project Implicit to participate in research on implicit bias. We used the Black–White implicit attitude test as the core tool to assess and provide feedback about bias. For this reason, we further constrained our sample to U.S. citizens because of the unique history and meaning of race in the United States, and because most Americans today believe in racial equality (see Charlesworth & Banaji, 2022). This approach is similar to other research examining reactions to bias feedback (Howell et al., 2013; Vitriol & Moskowitz, 2021; for an exception, see Howell & Ratliff, 2016), yet none so far have examined two interventions that have the virtue of simplicity. In the first intervention (Experiment 1), participants offer their opinion about the validity of a test they took before or after receiving the test result. We hypothesized that first affirming the validity of the test procedure would lead participants to accept a subsequently provided test result. In the second intervention (Experiments 2 and 3), we created two conditions. In the first condition, we presented a morally sensitive test (Black–White attitude) and elicited an evaluation of it. In the second condition, the same morally sensitive test was presented and evaluated, but after a morally innocuous test (Insect–Flower attitude) was first presented and evaluated. In all three experiments, we test whether the human desire to act consistently facilitated awareness and acceptance of bias. We additionally measured support for actions to minimize the effects of bias to test whether these interventions went beyond awareness and acceptance of bias to outcomes critical to achieving effective bias education.

Experiment 1

Overview and Design

In Experiment 1, data from three independent samples were collected and meta-analyzed. Each experiment adopted an identical procedure and design but utilized different measures, which improved slightly upon the measures used in the previous iteration of the data collection (see the online supplemental materials for all details of variations in measures).

Type of Test

The first independent variable varied (between subjects) the topic that was tested. In each data collection, two IATs on two different topics were administered, one to each half of the sample: a Black–White race IAT or an Insect–Flower IAT. Theoretical interest centered on the Black–White race IAT, which measures the strength of associations between the social groups Black–White and the attributes of good–bad. This test can produce an unflattering result about one's assumed race neutrality that commonly elicits resistance (Howell et al., 2015; Vitriol & Moskowitz, 2021). This occurs precisely because, in the United States, many White Americans genuinely experience a commitment to racial equality as well as a desire to be or appear nonprejudiced in the eyes of others (Plant & Devine, 1998; Bobo, 2001; O'Brien et al., 2010; Vitriol, 2016). For example, this test shows a White = good/Black = bad effect in over 70% of White and Asian participants (Morehouse & Banaji, 2024). Not surprisingly, these results often activate resistance to evidence about implicit bias, especially in those participants who believe themselves to be free of race bias (Moskowitz & Vitriol, 2021). A comparison IAT, one whose result is more likely to be accepted, was administered to the other half of the participants. An Insect–Flower IAT, which measures the association between the concepts Insect–Flower with the attributes good–bad, served as this comparison test. This test shows an association of Flower = good/Insect = bad in over 80% of participants (Banaji & Greenwald, 2013). Because these results are consistent with explicit preferences, it was not expected to produce resistance to accepting the test result.

Timing of Test Evaluation

The critical dependent variable was a measure of participants' evaluation of the IAT's validity and utility of a particular test result. The underlying assumption is that if the participant evaluates the validity and utility of the IAT (a measure we will refer to as assessment of IAT validity and utility [IAT-VU]) prior to receiving the test result, that evaluation of test validity and utility cannot be influenced by knowledge of the test result. In other words, there is no strong motivation to discredit the test before the feedback casts one in an unfavorable light. In the first evaluation of test validity and utility (prior to receiving the test results), a positive evaluation of the test would be difficult to contradict after receiving the test score, whether it is favorable or not to one's self-concept.

To measure how participants' test evaluation is influenced by the test result, we varied the timing of the evaluation of test validity and utility. This second independent variable, timing of test evaluation, had two levels. Half of the participants provided an evaluation of the test's validity and utility on the assessment IAT-VU both before and after receiving test score feedback. The other half of the participants provided an evaluation of the test's validity and utility on the assessment IAT-VU only after receiving test score feedback. We refer to participants in the former group as the before + after condition, and participants in the latter group as the after only condition.

All three data collections that constitute Experiment 1 employed a 2 (Type of IAT; Black–White [Critical] Test vs. Insect–Flower [Control] Test) × 2 (Timing of Test Evaluation; Before + After vs. After Only Test Result) between-subjects design. Participants were randomly assigned to complete one of two IATs—Black–White Race IAT or Insect–Flower IAT—and were then either assigned to the before + after evaluation or after only evaluation. Figure 1 provides a graphical representation of the procedural steps for participants across conditions. Below, we describe the characteristics of each sample and the combined sample used in Experiment 1. The online supplemental materials provide a complete description of the measures, analyses, and results for each sample in Experiment 1, separately.

Participants

Participants in all three samples that constitute Experiment 1 were independently recruited on the Project Implicit educational website (https://implicit.harvard.edu/implicit/). Because all three variations of Experiment 1 show the same pattern of results, we report the meta-analytic results for the combined sample, which included 2,025 White U.S. citizens ($M_{age} = 36.82$, $SD = 15.33$). Most participants had earned at least a bachelor's degree (59.53%) and self-identified as female (71.31%; male = 28.59%). The sample also skewed liberal in ideological self-placement ($M = 4.73$, $SD = 1.72$ on a 7-point scale ranging from 1 = *strongly conservative* to 7 = *strongly liberal*).

Figure 1
Experiment 1: Design and Procedure × Condition

	IAT (Black-White or Insect-Flower)	Assessment IAT-VU (Before) (pre-feedback of IAT Score)	Feedback on IAT Score	Assessment IAT-VU (After) (post-feedback of IAT Score)	IAT Feedback Acceptance	Bias Awareness
After Only	✓		✓	✓	✓	✓
Before + After	✓	✓	✓	✓	✓	✓

Note. IAT = Implicit Association Test; VU = validity and utility.

We conducted a power analysis for small and medium-sized effects based on this combined sample of White U.S. citizens. For a significant two-way interaction between the type of IAT and timing of test evaluation, we estimated that this experiment had at least 62% power to detect a Cohen's *f* of 0.10% and 99% power to detect a Cohen's *f* of 0.25. To detect mean-level differences between the before + after (vs. after only) condition within each IAT condition (Black–White IAT $n = 1{,}049$; Insect–Flower IAT $n = 976$), separately, we estimated that the experiment had at least 88% power to detect a Cohen's *d* of 0.2% and 99% power to a Cohen's *d* of 0.5 or higher within each IAT condition. Finally, sensitivity power analysis for the mediation analysis estimates showed at least 99% power to detect an indirect effect of Cohen's *d* of 0.2 or higher when comparing before + after (vs. after only) condition, collapsed across the IAT condition ($N = 2{,}025$).

Measures

The exact language used in the instructions, question stems, and items for each experiment are available in the online supplemental materials, along with the means, standard deviations, alphas, and intercorrelations of all measures separately for each sample.

IAT

The IAT is a computer-administered categorization task (Greenwald et al., 1998). Individuals are repeatedly presented with paired concepts and attributes, and faster response latencies indicate stronger implicit associations between the concept and attribute pairings. Half of the participants completed a Black–White attitude IAT, whereas the other half completed an Insect–Flower attitude IAT, each assessing the association between the concept categories (Black–White or Insect–Flower) with the attributes good and bad. In the Black–White IAT, participants were presented with pictures of White and Black men, and good (e.g., joy, wonderful) or bad (e.g., agony, hurt) words. In the Insect–Flower IAT, participants were presented with pictures of Insects and Flowers, and the same set of good or bad words.

The Insect–Flower IAT consists of two critical blocks: in one block, the labels "Flowers" and "Good" share the same response key, and "Insects" and "Bad" share another response key. A trial involves a picture of a stimulus appearing at the center of the screen (e.g., a picture of an Insect or Flower), which corresponds to one of the four labels, and the correct response key must be made before moving onto the next trial. In the other critical block, the instruction is reversed, so that the labels "Flowers" and "Bad" share the same response key and "Insects" and "Good" share the same response key. If participants have faster reaction times to the first block relative to the second block, this indicates a pro-Flower/anti-Insect bias. The magnitude of this difference is reflected in a participant's *D* score which is a measure of the size of the effect (see Greenwald et al., 2003). For the race IAT, the procedure is the same as for the Insect–Flower IAT, however, the categories "Insect" and "Flower," were replaced with social group categories, "White" and "Black." Furthermore, instead of using pictures of "Insect" and "Flower" as the stimuli, the race IAT uses pictures of faces as stimuli. To experience the test, readers can visit https://www.implicit.harvard.edu, use the link to the demo site (bottom left link), and choose any one of several other IATs. The site also contains answers to frequently asked questions.

Assessment IAT-VU

The same set of measures assessing the VU of the IAT was administered both before and after receiving test score feedback. We refer to the administration of assessment IAT-VU prior to or after feedback as assessment IAT-VU before and assessment IAT-VU after, respectively. We designed four items specifically for Experiment 1 and each was assessed on a 7-point scale (1 = *strongly disagree*, 2 = *moderately disagree*, 3 = *slightly disagree*, 4 = *neither agree nor disagree*, 5 = *slightly agree*, 6 = *moderately agree*, 7 = *strongly agree*). For example, the items in Experiment 1a included (a) "I believe the results of this test can be valid," (b) "This test can help us understand the way human minds works," (c) "This test can help us think about how humans interact with other groups," and (d) "This test measures thoughts of which we may be unaware." However, the language of the items and response scales varied across the three experiments (see the online supplemental materials).

Test Score Feedback Acceptance

To evaluate participants' acceptance of their test score feedback, participants were administered a single item designed specifically for Experiment 1, after receiving test score feedback and after completing the Assessment IAT-VU. "Do you agree with the results of the test?" was assessed on a 7-point scale (1 = *strongly disagree*, 2 = *moderately disagree*, 3 = *slightly disagree*, 4 = *neither agree nor disagree*, 5 = *slightly agree*, 6 = *moderately agree*, 7 = *strongly agree*). However, the language used for the item and response scales varied slightly across the three experiments (see the online supplemental materials).

Bias Awareness

Participants responded to Perry et al.'s (2015) Bias Awareness questionnaire, which was only administered after feedback on test

scores and after the aforementioned measures were administered. For this scale, participants responded to four items on a 7-point scale ranging from *strongly disagree* to *strongly agree*. Items included (a) "Even though I know it's not appropriate, I sometimes feel that I hold unconscious negative attitudes toward Black people," (b) "When interacting with Black, people I sometimes worry that I am unintentionally acting in a prejudiced way," (c) "Even though I like Black people, I still worry that I have unconscious biases toward Black people," and (d) " I never worry that I may be acting in a subtly prejudiced way toward Black people." Higher values were coded to indicate higher levels of bias awareness.

Transparency and Openness

All data exclusions (if any), all manipulations, and all measures in Experiment 1 are fully described. Data were analyzed using STATA, Version 17.0. Ethical approval was granted by the authors' Institutional Ethics Committee for all research activities reported in Experiment 1. The data and data syntax for the analysis are available at https://osf.io/7tfcs/?view_only=0819a809d2964aec8fe44ff998db601c.

Procedure

The procedure was identical across all data collections in Experiment 1, and consisted of the following in chronological order (a) a Black–White IAT or Insect–Flower IAT, (b) assessment IAT-VU, prior to feedback on test scores (for participants in the before + after but not after only group), (c) test score feedback, and (d) assessment IAT-VU (assessment IAT-VU after; all participants), test score feedback acceptance, and bias awareness. Figure 1 provides a graphical representation of the procedural steps for participants across conditions.

At the start of Experiment 1, participants reviewed a consent form. They were then randomly assigned to one of two IAT tests (Black–White race critical test vs. Insect–Flower comparison test). 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 one of the two IATs, half of the participants assigned to each test were placed in the before + after or after only test evaluation condition. In the before + after condition, participants completed the Assessment IAT-VU (described in detail above). Then, they received feedback on the results of their test. In the after only condition, participants' views on the test's validity and utility were not assessed at this stage. Rather they received feedback on the results of their IAT immediately after completing it.

Before receiving feedback on the results of the test, participants were provided the following preamble to their IAT scores:

> 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 / Insects and Flowers) with Good and Bad words.

Based on test scores, which were computed using the new scoring algorithm from Greenwald et al. (2009), all participants received one of seven possible forms of feedback about their performance on the IAT they completed. The feedback provided to participants in these studies is standard feedback given to all participants who complete an IAT at Project Implicit. Such feedback ranged from strong (vs. moderate vs. slight) automatic preference for (Black vs. White people/Insects vs. Flowers) to a strong (vs. moderate vs. slight) automatic preference for (White vs. Black people/Flowers vs. Insects), with no preference for (Black or White people/Insects or Flowers) as the midpoint.

After receiving feedback on the test results, all participants were provided with additional information intended to provide context for understanding the logic of the test and interpreting its results. The exact language of this information is provided in the online supplemental materials. Next, participants completed measures (described above) (a) assessing the validity and utility (assessment IAT-VU) of the IAT a second time (if in the before + after condition) or for the first time (if in the after only group), (b) a measure evaluating test score feedback acceptance, and (c) a measure of awareness of bias and concern about its influence on one's judgment and behavior. All participants were then debriefed and thanked for their time.

Results

The data from each of the three independent samples that constitute Experiment 1 showed highly similar patterns of results, but each lacked statistical power to reliably observe small effects when estimated separately. As such we present meta-analytic estimates, collapsing across the three data collections as the most reliable report of the data. Analyses of each sample are also reported separately in the online supplemental materials.

Table 1 reports the means, standard deviations, alphas, and intercorrelations of all measures, aggregated across all three experiments utilizing the same procedure and design, but with minor variations in the wording of the measures. The overall mean reported in Table 1 indicates that participants showed a pro-Flower/anti-Insect or pro-White/anti-Black implicit association in respective IAT conditions. The direction and magnitude of preference observed in each test are consistent with results observed in previous research (and are significantly different from zero, $ps < .001$). Notably, IAT results are not significantly correlated with bias awareness (Black–White D, $r = .03$, $p = .395$; Insect–Flower D, $r = .05$, $p = .129$). This finding indicates that performance on the measure of implicit cognition (IAT) is unrelated to one's awareness of bias and can be taken, in our view, as additional evidence of the dissociation between implicit and explicit cognition, although there are varied views on the meaning of such a pattern. Including IAT *D* scores as a covariate did not statistically nor substantively change our conclusions. In contrast, the explicit measure of bias awareness was positively and significantly correlated with other explicit measures, assessment IAT-VU both before and after test result feedback, and increased test score feedback acceptance (assessment of IAT-VU before, $r = .46$, $p < .001$; assessment of IAT-VU after, $r = .53$, $p < .001$; test score feedback acceptance, $r = .45$, $p < .001$). That is, more favorable assessments of IAT validity and utility and increased acceptance of its results, not performance on the IAT itself, predicted increased levels of bias awareness.

To examine the effects of the independent variables (type of test and timing of test evaluation), we conducted a mini meta-analysis

Table 1
Mean, Standard Deviation, and Intercorrelation of Measures Administered in Experiment 1

Variables	*M*	*SD*	Alpha [95% CI]	Insect–Flower IAT D	Black–White IAT D	IAT-VU before	IAT-VU after	FB acceptance	Bias aware
Insect–Flower IAT D	0.60	0.38	—	—					
Black–White IAT D	0.41	0.42	—	—	—				
IAT-VU before	4.39	1.49	0.87 [0.84, 0.90]	.03	−.04	—			
IAT-VU after	4.39	1.55	0.91 [0.89, 0.93]	.10**	−.10**	.84**	—		
FB acceptance	4.29	1.64	—	.17**	−.18**	.73**	.88**	—	
Bias awareness	3.81	1.53	0.85 [0.83, 0.86]	.05	.03	.46**	.53**	.45**	—

Note. Higher values correspond with pro-Flower/anti-Insect, pro-White/anti-Black, more favorable attitudes of test validity and utility, increased acceptance of test results feedback, and awareness of bias. The means, standard deviations, and intercorrelations reported here are aggregated across all three samples in Experiment 1; see the online supplemental materials for information about each sample. The information contained in these tables is based on all available observations. IAT D = results of the Implicit Association Test; IAT-VU before = assessment of test validity and utility prior to feedback on test results; IAT-VU after = assessment of test validity and utility after feedback on test results; FB acceptance = acceptance of the results of the IAT; bias awareness = awareness of and concern about one's own implicit racial bias; CI = confidence interval; IAT = Implicit Association Test; VU = validity and utility.
** $p < .01$.

(Goh et al., 2016; also referred to as a mega-analysis, see Eisenhauer, 2021) for the main and (when appropriate) interaction effects across the three experiments. This analysis was done by pooling the raw data and then estimating a multilevel model with maximum likelihood estimation and with experiment submitted as a random-intercept term. Effect sizes were estimated using inverse variance weighting (Goh et al., 2016).

Does Assessment IAT-VU (Before) Vary Across IAT Conditions, Prior to Feedback on Test Results?

We first examined differences in Assessment IAT-VU before (assessment IAT-VU before) test score feedback, between the two IAT conditions (Black–White and Insect–Flower attitude test). Doing so allows us to gauge differences in perceptions of the two IATs before feedback on the test results is provided. Assessment IAT-VU (before) did not differ between the two IAT conditions ($Z = -0.02$, $b = -0.002$, $SE = 0.04$, 95% CI [−0.07, 0.07], $p = .98$). Together, these results indicate that, prior to feedback on the results of the IAT, participants did not differ in their perceptions of the Black–White compared to the Insect–Flower IAT.

Did Assessment IAT-VU (After) and Test Score Feedback Acceptance Vary as a Function of Timing of Test Evaluation and Type of IAT?

Next, we examine the effect of our two independent variables (timing of test evaluation, type of IAT) on the two critical dependent variables (assessment IAT-VU, test score feedback acceptance), which were administered after participants received feedback on the test results. We anticipated that, compared to the after only group, the before + after group (i.e., completing the assessment IAT-VU prior to feedback on test scores) would demonstrate more favorable assessments of the IAT (assessment IAT-VU after) and increased levels of test score feedback acceptance. We also reasoned that this effect would be particularly likely for participants in the Black–White IAT, but not the Insect–Flower IAT, condition. In contrast, no such motivation is present for participants in the Insect–Flower IAT condition, so we do not anticipate an effect of before + after (vs. after only) on postfeedback of test results assessment IAT-VU (after) or levels of test score feedback acceptance. Figure 2 graphically represents the mean of each dependent variable across conditions.

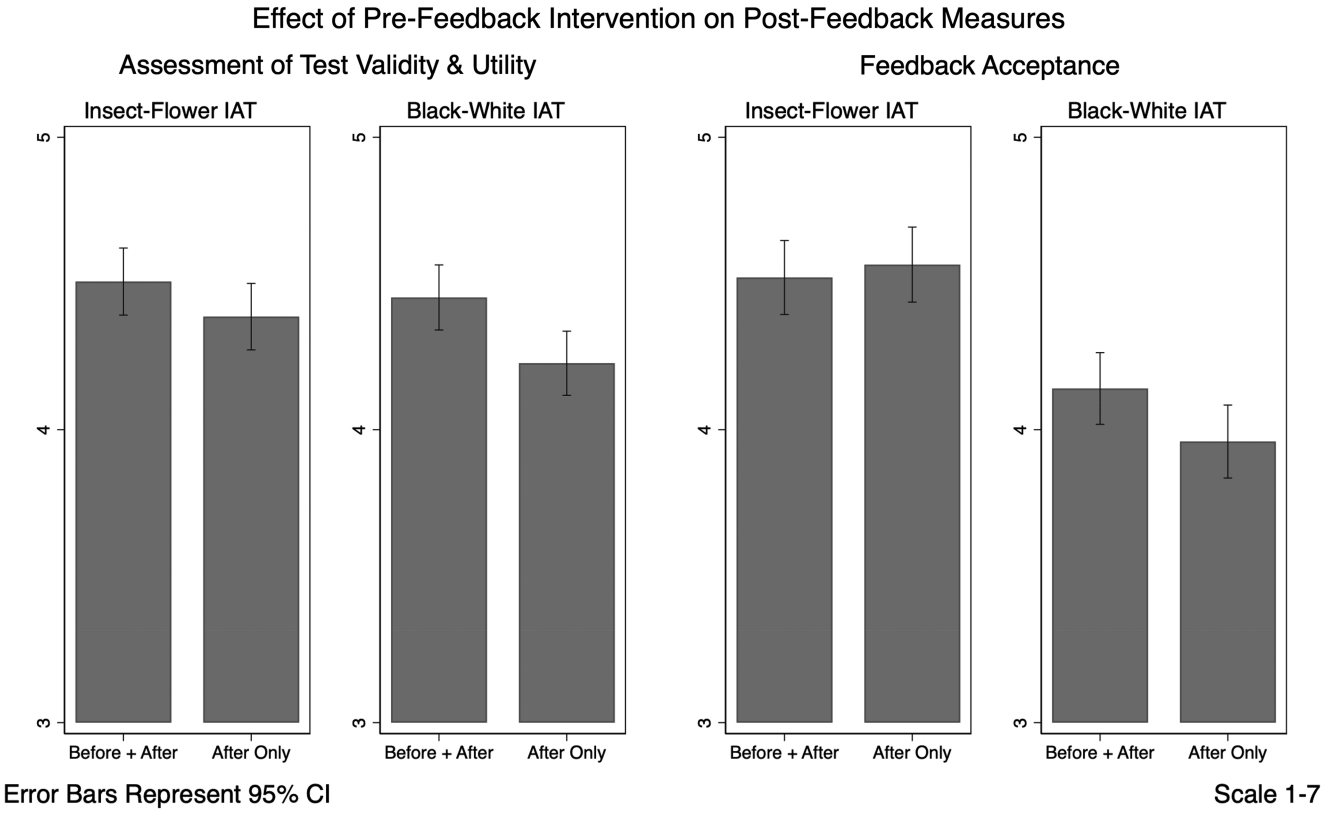
Assessment IAT-VU (After)

The result of the random-intercept model for assessment IAT-VU (after) did not yield a significant interaction between timing of test evaluation and type of IAT ($Z = -0.92$, $b = -0.03$, $SE = 0.03$, 95% CI [−0.08, 0.03], $p = .36$), suggesting that the effect of timing of test evaluation did not differ significantly between the two IAT conditions. Furthermore, we did not observe a significant effect for the Type of IAT (Black–White IAT, $M = 4.33$, $SD = 1.55$; Insect–Flower IAT, $M = 4.46$, $SD = 1.55$; $Z = -1.86$, $b = -0.05$, $SE = 0.03$, 95% CI [−0.11, 0.003], $p = .063$, Cohen's $d = 0.11$). While the effects were not statistically significant, the pattern nonetheless suggests that participants reported more favorable assessments of the Insect–Flower IAT than the Black–White IAT. More importantly, we observed a significant main effect of timing of test evaluation on assessment IAT-VU (after) (after only, $M = 4.31$, $SD = 1.55$; before + after, $M = 4.47$, $SD = 1.55$; $Z = -3.00$, $b = -0.09$, $SE = 0.03$, 95% CI [−0.14, −0.03], $p = .003$, Cohen's $d = 0.13$). That is, these findings indicate that before + after (vs. after only) led, after test result feedback, to more favorable assessments of the IAT.

We did not observe an interaction between timing of test evaluation and type of IAT on assessment IAT-VU (after). Nonetheless, we expected a priori that the effect of timing of test evaluation would be larger in the morally challenging Black–White IAT condition, compared to the more innocuous Insect–Flower IAT condition. For this reason, we explored differences between the before + after and after only groups on assessment IAT-VU (after), separately in the Black–White and Insect–Flower IAT conditions. For participants in the Black–White IAT condition, we obtain significant effects of before + after (vs. after only) on assessment IAT-VU (after) ($Z = -2.70$, $b = -0.12$, $SE = 0.04$, 95% CI [−0.19, −0.03], $p = .007$, Cohen's $d = 0.16$; after only, $M = 4.22$, $SD = 1.53$; before + after, $M = 4.44$, $SD = 1.58$). However, for participants in the Insect–Flower IAT condition, we did not observe significant effects of before + after (vs. after only) on assessment IAT-VU (after) ($Z = -1.58$, $b = -0.06$, $SE = 0.04$, 95% CI [−0.14, 0.02], $p = .114$; after only, $M = 4.42$, $SD = 1.56$; before + after, $M = 4.50$, $SD = 1.53$). Thus,

Figure 2

Between-Subject Comparison of Mean-Level Responses × Condition, Experiment 1



Note. Before + after = participants evaluated test validity and utility prior to feedback; after only = participants did not evaluate test validity and utility prior to feedback; higher values = more favorable attitudes and increased acceptance; IAT = Implicit Association Test; CI = confidence interval.

the effect of timing of test evaluation on assessment IAT-VU (after) was significantly different from zero in the Black–White IAT, but not the Insect–Flower IAT condition. Nonetheless, we must caution against interpreting the results of this post hoc analysis as strong evidence for a differential effect, across the two IAT conditions, of timing of test evaluation on assessment IAT-VU (after), given that the interaction term was not statistically significant (see Gelman & Stern, 2006).

Test Score Feedback Acceptance

The result of the random-intercept model for test score feedback acceptance did not yield a significant interaction between timing of test evaluation and type of IAT ($Z = -1.76$, $b = -0.06$, $SE = 0.03$, 95% CI [−0.12, 0.01], $p = .078$). Furthermore, we did not observe a significant main effect for timing of test evaluation on feedback acceptance ($Z = -1.07$, $b = -0.03$, $SE = 0.03$, 95% CI [−0.10, 0.03], $p = .286$). However, we did observe a significant main effect for the type of IAT (Black–White IAT, $M = 4.04$, $SD = 1.70$; Insect–Flower IAT, $M = 4.04$, $SD = 1.70$; $Z = -7.68$, $b = -0.25$, $SE = 0.03$, 95% CI [−0.12, 0.006], $p < .001$, Cohen's $d = 0.32$). These findings indicate that test score feedback acceptance was higher for participants in the Insect–Flower, compared to the Black–White IAT, condition.

While the interaction was not statistically significant, we nonetheless expected a priori that the effect of timing of test evaluation would be larger in the morally challenging Black–White IAT condition, compared to the more innocuous Insect–Flower IAT condition. This expectation for this reason, we explored differences between the before + after and after only groups on assessment IAT-VU (after), separately in the Black–White and Insect–Flower IAT conditions. To decompose this interaction, we tested for differences between the before + after and after only groups on test score feedback acceptance separately in the Black–White and Insect–Flower IAT conditions. For participants in the Black–White IAT condition, we did not observe significant effects of before + after (vs. after only) on test score feedback acceptance ($Z = -1.90$, $b = -0.09$, $SE = 0.05$, 95% CI [−0.19, 0.003], $p = .058$, Cohen's $d = 0.11$; after only, $M = 3.96$, $SD = 1.67$; before + after, $M = 4.11$, $SD = 1.73$). Nonetheless, the effects trended in the hypothesized direction, suggesting that timing of test evaluation increased acceptance of the results of the Black–White IAT. However, for participants in the Insect–Flower IAT condition, we did not observe significant effects of before + after (vs. after only) on test score feedback acceptance ($Z = 0.56$, $b = 0.03$, $SE = 0.04$, 95% CI [−0.06, −0.11], $p = .575$; after only, $M = 4.51$, $SD = 1.53$; before + after, $M = 4.61$, $SD = 1.51$). Overall, we find that before + after (vs. after only) led, after test result feedback, to increased acceptance of its results in a socially

sensitive domain (i.e., Black–White IAT), but had no effect on acceptance of the results in a domain in which the feedback is unlikely to threaten importantly held values or social norms (i.e., Insect–Flower IAT). We again caution against interpreting the results of this post hoc analysis as evidence for a difference between the two IAT conditions of timing of test evaluation on assessment IAT-VU (after), given that the interaction term was not statistically significant (see Gelman & Stern, 2006).

Does Timing of Test Evaluation Indirectly Effect Bias Awareness?

Finally, we examined whether time of test evaluation would indirectly increase bias awareness by increasing more favorable assessment IAT-VU (after) and test score feedback acceptance. To test this, we examined whether postfeedback assessment IAT-VU (after) or test score feedback acceptance mediates the effects of before + after (vs. after only) on bias awareness, collapsed across IAT conditions. In order to account for clustering of responses within experiments in this mediation analysis, the indirect effect was computed based on the product-of-coefficient approach, using the multilevel mediation analysis command available in STATA that was adapted from Krull and MacKinnon (2001). Subsequently, a bootstrap analysis was performed following the recommendation by Preacher and Hayes (2004) with 5,000 resampled data sets. Bootstrapping estimates the indirect effect on each resampled data set based on the null hypothesis that the indirect effect is not different from zero. For all analyses below, we reject the null hypothesis if the confidence interval does not include zero (Preacher & Hayes, 2004).

Results indicate that, with condition submitted as an independent variable (0 = *before + after*, 1 = *after only*) and post feedback assessment IAT-VU (after) submitted as a mediator, the indirect and total effects of the before + after (vs. after only) on bias awareness was significant (indirect effect, $b = -0.03$, $SE = 0.02$, 95% CI [−0.06, −0.003], $p = .030$; total effect, $b = -0.05$, $SE = 0.02$, 95% CI [−0.09, −0.01], $p = .009$). However, we did not observe a significant direct effect ($b = -0.02$, $SE = 0.03$, 95% CI [−0.07, 0.04], $p = .58$). That is, by increasing favorable assessment IAT-VU (after), the timing of test evaluation indirectly increased bias awareness. Similarly, with condition submitted as an independent variable (0 = *before + after*, 1 = *after only*) and test score feedback acceptance submitted as a mediator, the indirect effects of the before + after (vs. after only) on bias awareness ($b = -0.11$, $SE = 0.04$, 95% CI [−0.19, −0.03], $p = .008$) obtained significance, but not the direct effect ($b = 0.05$, $SE = 0.04$, 95% CI [−0.02, 0.12], $p = .18$) nor total effect ($b = -0.06$, $SE = 0.08$, 95% CI [−0.21, 0.09], $p = .43$). These findings again indicate that the before + after (vs. after only) condition indirectly increased bias awareness by increasing test score feedback acceptance.

Overall, this pattern of results indicates that the effect of timing of test evaluation on bias awareness is fully mediated by its effects on assessments of the IAT and test score feedback acceptance.

Experiment 2

Overview and Design

The results of Experiment 1 demonstrated that evaluating the validity and utility of a test of implicit attitudes, and its results, before receiving test result feedback, can promote increased bias awareness. This increased awareness emerged in conditions that motivated more favorable evaluations of the test and increased acceptance of its results (after the test score feedback was provided), although the latter effect approached but did not reach conventional levels of statistical significance. From this, we infer that creating conditions for objectivity and consistency (here, created in the before + after condition) can elicit greater awareness of bias. Furthermore, we observed some evidence that these effects are specific to a socially sensitive domain (i.e., Black–White race IAT), in which the test result feedback can threaten moral values and social norms. Experiment 2 was designed to explore the same idea of creating conditions that promote acceptance or rejection of test validity and utility, using a different experimental paradigm.

In Experiment 2, we created a second independent path to creating the conditions that promote or dampen objectivity and consistency in assessing one's implicit bias. We gave half the participants a first test whose result would not be morally challenging and thus more easily accepted (Insect–Flower IAT and result), followed by the Black–White IAT. The rationale was simple. Administering the Insect–Flower test first and obtaining an evaluation of test validity and utility would make it difficult to resist acceptance of the Black–White test result when it is subsequently presented. Theoretical interest centered on perceptions of the Black–White IAT and acceptance of its results among participants who had first completed the Insect–Flower IAT compared to those who only completed the Black–White IAT.

In addition to varying whether or not participants first completed the Insect–Flower IAT prior to the Black–White IAT, we also introduced a third condition in Experiment 2. In this additional condition, we provided participants with information that highlights the similarities in the underlying logic and meaning of the two tests (see below for the exact language used in this condition). This information also impressed upon participants the threats to objectivity and consistency in rejecting the results of one test while accepting the results of another test based on the same logic (i.e., that the process by which the IAT computes a result is the same for the Insect–Flower test as well as the Black–White test). We anticipated that explicitly emphasizing the shared features of the Insect–Flower and Black–White IATs would help to reinforce their logical connection, further weakening resistance to acknowledging implicit bias (see Hughes et al., 2020). In addition to emphasizing the logical connections between the two IATs, this information also emphasized the role of cultural learning in the acquisition of implicit attitudes, which other research suggests may help increase acceptance of IAT feedback (see Uhlmann & Nosek, 2012; Vitriol & Moskowitz, 2021).

Accordingly, Experiment 2 utilized the measures from Experiment 1b, and employed three conditions: (a) Black–White IAT only (B/W test only), which represents the condition under which we expect resistance to be highest; (b) Insect–Flower IAT followed by Black–White IAT (I/F then B/W test); and (c) Insect–Flower IAT followed by Black–White IAT, in which additional information about the IAT was provided to explicitly connect the logic underlying the methodology of both tests (I/F then B/W test + connect). By comparing the latter two conditions, we are able to explore the impact of emphasizing the shared methodology, logic, and meaning of the IATs. Figure 3 provides a graphical representation of the procedural steps for participants across conditions.

Figure 3

Experiment 2 Procedure × Condition

	Insect-Flower IAT					Black-White IAT				Post-IAT Measures	
Condition	IAT	Feedback on IAT Score	Assessment IAT-VU	IAT Feedback Acceptance	Intervening Instructions	IAT	Feedback on IAT Score	Assessment IAT-VU	IAT Feedback Acceptance	Bias Awareness	Act Against Bias
B/W Test Only						✓	✓	✓	✓	✓	✓
I/F then B/W Test	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
I/F then B/W Connect	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note. B/W test only = participants only completed the Black–White IAT; I/F then B/W test = participants completed the Insect–Flower IAT first, followed by the Black–White IAT second; I/F then B/W + connect = participants completed the Insect–Flower IAT first, following by the Black–White IAT second, with additional information provided to transition between IATs; IAT = Implicit Association Test; VU = validity and utility; B/W = Black/White; I/F = Insect/Flower.

Participants

In Experiment 2, we recruited 809 White U.S. citizens as volunteers from the Project Implicit educational website (https://implicit.harvard.edu/implicit/) to participate in Experiment 2. Of these, 63 were excluded from analyses for failing to complete the first (or only) IAT to which they were assigned. A total of 746 White U.S. citizens completed the full survey ($M_{age} = 44.30$, $SD = 15.21$). Most participants had earned at least a bachelor's degree (59.5%) and self-identified as female (70.24%; male = 29.09%). The sample also skewed liberal in ideological self-placement ($M = 5.31$, $SD = 1.65$ on a 7-point scale ranging from 1 = *strongly conservative* to 7 = *strongly liberal*). We conducted a sensitivity analysis for small and medium-sized effects based on this sample of White U.S. citizens.

With the current sample size, to detect mean level differences between the three conditions, our estimate is that this experiment had at least 68% power to detect a Cohen's *d* of 0.2% and 99% power to detect a Cohen's *d* of 0.5 or higher. The sensitivity power analysis for the mediation analysis estimates that we had at least 89% power to detect an indirect effect of Cohen's *d* of 0.2 or higher when comparing I/F then B/W + connect to B/W test only condition ($N = 500$).

Measures

Means, standard deviations, alphas, and intercorrelations of all measures are available in Table 2. The exact language used in the instructions, question stems, and items are available in the online supplemental materials. All of the measures used in Experiment 2 are identical to Experiment 1b (see the online supplemental materials), with the only exceptions or additions described below.

Assessment IAT-VU

We revised one item to improve clarity. The original Experiment 1b item was "This test can help us think about how humans interact with members of other groups." For Experiment 2, this item was revised as follows: "This test can help us think about human behavior." All four items were again assessed on a 7-point scale (1 = *strongly disagree*, 2 = *moderately disagree*, 3 = *slightly disagree*, 4 = *neither agree nor disagree*, 5 = *slightly agree*, 6 = *moderately agree*, 7 = *strongly agree*). Higher values represent more favorable attitudes toward the test.

Support for Actions to Mitigate Consequences of Bias

Items to evaluate support for actions to minimize the consequences of bias were assessed in Experiment 2. Participants were administered four items on a 7-point scale (1 = *strongly disagree*, 2 = *moderately disagree*, 3 = *slightly disagree*, 4 = *neither agree nor disagree*, 5 = *slightly agree*, 6 = *moderately agree*, 7 = *strongly agree*). These items include (a) "Given that racial bias can be invisible to us, it is important to support policies that reduce such biases," (b) "Being aware of one's racial bias and doing nothing about it isn't very smart," (c) "People make a big deal about racial bias, even though it hardly exists," and (d) "Affirmative action is needed to balance out the consequences of bias." Higher values represent higher levels of support for actions to combat the consequences of bias.

Transparency and Openness

All data exclusions (if any), all manipulations, and all measures in Experiment 2 are fully described. Data were analyzed using STATA, Version 17.0. Ethical approval was granted by the authors' Institutional Ethics Committee for all research activities reported in Experiment 2. The data and data syntax for the analysis are available at https://osf.io/7tfcs/?view_only=0819a809d2964aec8fe44ff998db601c.

Procedure

Experiment 2 included three conditions. For participants in the first condition (B/W test only), the procedure consisted of, in chronological order: (a) an Insect–Flower IAT, (b) feedback about the results of the Insect–Flower IAT, and (c) measures of assessment IAT-VU and test score feedback acceptance for the Insect–Flower IAT. For participants in the second condition (I/F then B/W test), the procedure consisted of, in chronological order: (a) an Insect–Flower IAT, feedback about the results of the Insect–Flower IAT, and measures of assessment IAT-VU and test score feedback acceptance for the Insect–Flower IAT, (b) a Black–White IAT, (c) feedback about the results of the Black–White IAT, and (d) measures of Assessment IAT-VU and test score feedback acceptance for the Black–White IAT. For participants in the third condition (I/F then B/W test + connect), the procedure was identical to that of participants in the second condition. The notable exception to this is the inclusion of information that connected the logic underlying the

Table 2
Mean, Standard Deviation, and Intercorrelation of Measures Administered in Experiment 2

Variables	*M*	*SD*	Alpha [95% CI]	Bug IAT D	Race IAT D	Bug IAT-VU	Race IAT-VU	Race FB acceptance	Bug FB acceptance	Act against bias	Bias aware
Insect–Flower IAT D	0.66	0.38	—	—							
Black–White IAT D	0.28	0.45	—	.15**	—						
Race IAT-VU	5.35	1.28	.88 [0.85, 0.91]	.01	−.08*	—					
Bug IAT-VU	5.43	1.23	.92 [0.90, 0.93]	.11*	.00	.63**	—				
Race FB acceptance	4.86	1.57	—	−.02	−.18**	.66**	.36**	—			
Bug FB acceptance	5.28	1.78	—	.10*	−.05	.40**	.61**	.33**	—		
Act against bias	6.24	0.93	.69 [0.62, 0.74]	−.05	−.17**	.36**	.23**	.20**	.07**	—	
Bias awareness	4.82	1.38	.80 [0.77, 0.83]	−.06	.01	.38**	.21**	.26**	.12**	.27**	—

Note. All explicit measures are assessed on a 1–7 scale. Higher values correspond with pro-Flower/anti-Insect, pro-White/anti-Black, more favorable assessments of test validity and utility, increased willingness to support action against bias, and increased acceptance of bias feedback and awareness of bias. The information contained in these tables is based on all available observations. IAT D = results of the Implicit Association Test; IAT-VU = assessment of test validity and utility, for each IAT; FB acceptance = acceptance of the results of the IAT; bias awareness = awareness of and concern about one's own implicit racial bias; act against bias = willingness to support actions that combat the consequences of bias. CI = confidence interval; IAT = Implicit Association Test; VU = validity and utility.
* $p < .05$. ** $p < .01$.

methodology of both IATs and emphasized the role of cultural learning in the acquisition of implicit attitudes, which was provided before the start of the Black–White IAT (see Figure 3).

At the start of Experiment 2, participants reviewed a consent form and were then randomly assigned to condition. Before completing each IAT, all participants received the same pre-IAT instructions as in Experiment 1. After each IAT, participants were provided feedback about the results of the IAT in the same way as for Experiment 1. Among participants who completed both the Insect–Flower IAT and the Black–White IAT, half of participants (I/F then B/W + connect) were also provided the following information during the intervening period between IATs:

> In this next phase of the experiment you'll complete a test that will replace Flowers and Insects with faces of Black and White people. Otherwise, the test is identical. It's interesting to us that although people freely accept that the first test provides a true picture of their flower/insect attitudes, they are not able to accept the same with the race test, especially if it shows association of Black with bad. Being scientists, we cannot do that: we have to be rational and accept both results or neither. We see the results of these tests as cultural learning in both cases. People learn what our culture teaches us and it gets into our minds, whether it's about flowers, cats, cars, or race. Please continue.

Participants then completed measures that assessed perceived test validity and utility of the Black–White IAT (assessment IAT-VU) and test score feedback acceptance. After the Black–White IAT, in addition to completing these measures, participants also completed measures evaluating bias awareness and support for actions to mitigate the consequences of bias. All participants were then debriefed and thanked for their time.

Results

Table 2 reports the means, standard deviations, alphas, and intercorrelations of all measures. The overall mean reported in Table 2 indicates that, replicating the results of Experiment 1, participants showed a pro-Flower or pro-White implicit association in respective IAT conditions. The direction and magnitude of preference observed in each test is consistent with results observed in previous research and in Experiment 1 (and is significantly different from zero, $ps < .001$). We again observe that IAT results are not significantly correlated with bias awareness (Black–White D, $r = .01$, $p = .77$; Insect–Flower D, $r = -.06$, $p = .19$). This finding indicates that how participants performed on the IAT does not translate to increased awareness of bias. Consistent with Experiment 1, including IAT *D* scores as a covariate in the analysis did not statistically change the conclusions. In contrast, increased levels of bias awareness were positively and significantly correlated with more favorable assessment IAT-VU for both IATs and increased test score feedback acceptance (Black–White IAT assessment of IAT-VU, $r = .21$, $p < .001$, Black–White IAT feedback acceptance, $r = .26$, $p < .001$; Insect–Flower IAT Assessment of IAT-VU, $r = .27$, $p < .001$, Insect–Flower IAT feedback acceptance $r = .12$, $p = .027$). Additionally, increased levels of bias awareness correlated with increased support for actions against bias ($r = .27$, $p < .001$). That is, more favorable assessments of the IAT and increased acceptance of its results, not performance on the IAT itself, predict increased levels of bias awareness, which is significantly associated with increased support for actions against bias.

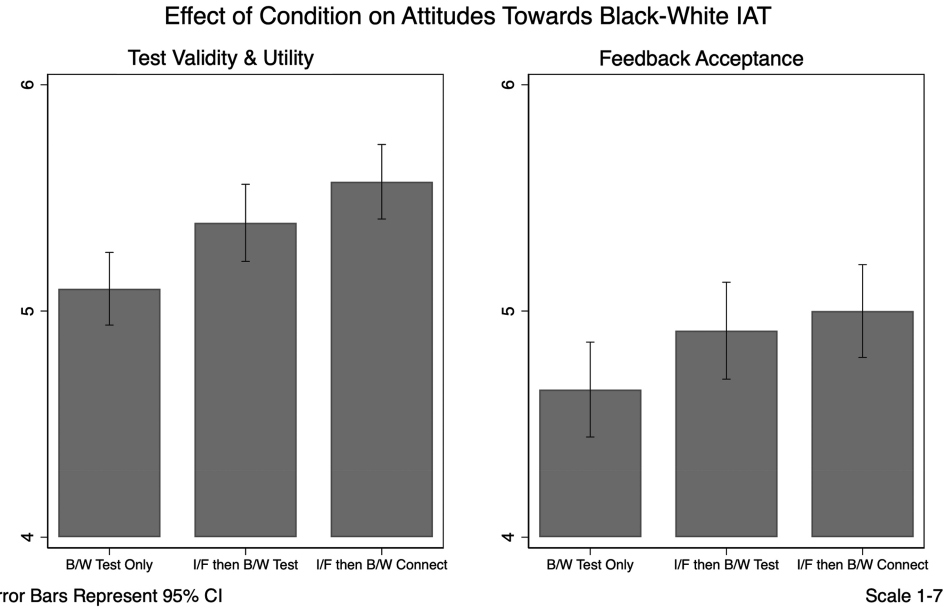
To examine the effects of our independent variable in Experiment 2, we conducted a one-way between-subjects ANOVA and, following a significant omnibus test, used Duncan's method for post hoc analyses to estimate pairwise comparisons.

Does Assessment IAT-VU and Test Score Feedback Acceptance Vary Across Conditions?

First, we examined differences across conditions for the two critical dependent variables (assessment IAT-VU, test score feedback acceptance), which were administered after participants completed the Black–White IAT. We anticipated that, compared to the Black–White only condition, the two conditions in which participants completed both the Insect–Flower IAT and Black–White IAT (i.e., Insect–Flower IAT first, Black–White IAT second, with or without a statement that connects the underlying logic of both IATs), would demonstrate more favorable assessments of the Black–White IAT (assessment IAT-VU) and increased levels of Black–White test score feedback acceptance. Figure 4 graphically represents the mean of each dependent variable across condition.

Figure 4

Between-Subject Comparison of Mean-Level Responses × Condition, Experiment 2



Note. B/W test only = participants only completed the Black–White IAT. I/F then B/W test = participants first completed the Insect–Flower IAT before the Black–White IAT. I/F then B/W + connect = participants first completed the Insect–Flower IAT before the Black–White IAT, with information that explicitly connects the logic underlying the methodology of both tests. Higher values = more favorable attitudes and increased willingness to accept feedback. IAT = Implicit Association Test; B/W = Black/White; I/F = Insect/Flower.

Assessment IAT-VU

We observe significant differences in assessment IAT-VU for the Black–White IAT, assessment IAT-VU, $F(2, 681) = 8.30$, $p < .001$, across conditions. Post hoc analyses using Duncan's method indicated that attitudes toward the Black–White IAT were less positive in B/W test only ($M = 5.10$, $SD = 1.32$), compared to I/F then B/W test ($M = 5.39$, $SD = 1.19$; 95% CI for $M_{\text{difference}}$ [0.06, 0.53], $p = .015$, Cohen's $d = .23$) and I/F then B/W test + connect ($M = 5.57$, $SD = 1.29$; 95% CI for $M_{\text{difference}}$ [0.23, 0.72], $p < .001$, Cohen's $d = 0.36$). The two conditions in which participants completed both the Insect–Flower IAT and Black–White IAT did not significantly differ for postfeedback test attitudes (95% CI for $M_{\text{difference}}$ [−0.06, 0.42], $p = .13$). These results demonstrate that participants first completing, receiving, and evaluating feedback from an Insect–Flower IAT test subsequently increases positive evaluation of the Black–White IAT.

Test Score Feedback Acceptance

We observe a similar pattern of results for Black–White IAT test score feedback acceptance, but this effect did not reach conventional levels of statistical significance across conditions, $F(2, 646) = 2.90$, $p = .056$. Post hoc analyses using Duncan's method indicated that test score feedback acceptance for the Black–White IAT was lower in B/W test only ($M = 4.66$, $SD = 1.62$), compared to I/F then B/W test ($M = 4.91$, $SD = 1.49$; 95% CI for $M_{\text{difference}}$ [−0.04, 0.56], $p = .089$, Cohen's $d = 0.17$) and I/F then B/W + connect ($M = 5.00$, $SD = 1.59$; 95% CI for $M_{\text{difference}}$ [0.04, 0.66], $p = .027$, Cohen's $d = 0.21$), although this effect again did not reach conventional levels of statistical significance. The two conditions in which participants completed both the Insect–Flower IAT and Black–White IAT also did not significantly differ for test score feedback acceptance (95% CI for $M_{\text{difference}}$ [−0.21, 0.38], $p = .565$). Thus, we did not observe strong evidence to suggest that first completing, receiving, and evaluating feedback from an Insect–Flower test subsequently increased acceptance of the result of the Black–White test.

Are There Indirect Effects on Bias Awareness and Support for Actions to Mitigate Bias?

Finally, we examined whether I/F then B/W test or I/F then B/W + connect (i.e., completed the Insect–Flower IAT first, followed by a Black–White IAT), compared to B/W test only (i.e., only completed a Black–White IAT), would indirectly increase bias awareness and support for actions to mitigate bias, by increasing more favorable assessment IAT-VU and test score feedback acceptance. To test this, we again use the bootstrap-based method recommended by Preacher and Hayes (2004), in which 5,000 bootstrap-replications were used to estimate confidence intervals, with assessment IAT-VU or test score feedback acceptance for the Black–White IAT submitted as mediators in separate analyses. To simplify this analysis, and because the effect sizes for the analyses described above were larger in I/F then B/W + connect (vs. I/F then B/W test) we compare the B/W test only condition to I/F then B/W + connect condition. Nonetheless the observed effects are similar when the comparison is between B/W test only and I/F then B/W test.

Bias Awareness

With condition submitted as an independent variable (0 = I/F then B/W + connect, 1 = B/W test only) and Assessment IAT-VU submitted as a mediator, the indirect effect of I/F then B/W + connect (vs. B/W test only) condition obtained significance on bias awareness ($b = -0.16$, $SE = 0.05$, 95% CI [−0.26, −0.07], $p = .001$). The direct effect of I/F then B/W + connect (vs. B/W test only) condition was not significant ($b = 0.03$, $SE = 0.12$, 95% CI [−0.21, 0.27], $p = .82$), nor was the total effect ($b = -0.13$, $SE = 0.13$, 95% CI [−0.38, 0.12], $p = .303$). That is, by increasing favorable assessment IAT-VU, the I/F then B/W + connect (vs. B/W test only) condition indirectly increased bias awareness. Similarly, with test score feedback acceptance submitted as a mediator, the indirect effect of I/F then B/W + connect (vs. B/W test only) condition on bias awareness approached but did not reach conventional levels of statistical significance ($b = -0.06$, $SE = 0.03$, 95% CI [−0.13, 0.002], $p = .056$). However, the direct effect of I/F then B/W + connect (vs. B/W test only) was not significant ($b = -0.10$, $SE = 0.13$, 95% CI [−0.36, 0.16], $p = .42$), nor was the total effect ($b = -0.17$, $SE = 0.13$, 95% CI [−0.42, 0.09], $p = .20$). Thus, we did not observe strong evidence for an indirect effect of the experimental manipulation on bias awareness through increased test score feedback acceptance.

Support for Actions to Mitigate Consequences of Bias

With condition submitted as an independent variable (0 = I/F then B/W + connect, 1 = B/W test only) and Assessment IAT-VU submitted as a mediator, the indirect effect of I/F then B/W + connect (vs. B/W test only) condition obtained significance on support for actions to mitigate the consequences of bias ($b = -0.12$, $SE = 0.04$, 95% CI [−0.19, −0.05], $p = .001$). The direct effect of I/F then B/W + connect (vs. B/W test only) condition was not significant ($b = 0.03$, $SE = 0.09$, 95% CI [−0.14, 0.20], $p = .726$), nor was the total effect ($b = -0.09$, $SE = 0.09$, 95% CI [−0.26, 0.08], $p = .302$). That is, by increasing favorable assessment IAT-VU, the I/F then B/W + connect (vs. B/W test only) condition indirectly increased support for actions to combat the consequences of bias. With test score feedback acceptance submitted as a mediator, the indirect effect of I/F then B/W connect (vs. B/W test only) condition on support for actions to mitigate the consequences of bias approached but did not reach statistical significance ($b = -0.04$, $SE = 0.02$, 95% CI [−0.08, 0.001], $p = .053$). However, the direct effect of I/F then B/W + connect (vs. B/W test only) was not significant ($b = -0.03$, $SE = 0.09$, 95% CI [−0.20, 0.15], $p = .42$), nor was the total effect ($b = -0.07$, $SE = 0.09$, 95% CI [−0.25, 0.11], $p = .439$). These findings do not provide strong evidence to suggest that I/F then B/W + connect (vs. B/W test only) indirectly increased support for actions to mitigate the consequences of bias by increasing test score feedback acceptance. However, the effect of the experimental manipulation on support for actions to mitigate bias is fully mediated by its effects on assessments of the IAT.

Experiment 3

Overview and Design

In Experiment 2, we found that participants who first completed and evaluated feedback from an Insect–Flower IAT test reported more positive attitudes following completion of a more socially sensitive test (i.e., Black–White IAT). As a result of this increase in positive attitudes, participants reported more awareness of bias and were more supportive of actions to mitigate the consequences of bias. We reasoned that participants' evaluating the Insect–Flower IAT likely constrained the motivation to ignore or dismiss evidence of bias following completion of the Black–White IAT. In other words, it provided the impetus for consistency in responding. However, Experiment 2 could not rule out the possibility that completing any IAT first would subsequently increase positive attitudes and test result acceptance of any subsequent IATs.

Experiment 3 was designed to replicate and extend the results of Experiment 2 by examining whether or not completing the Black–White IAT before the Insect–Flower IAT would also subsequently increase positive attitudes and test result acceptance of the Insect–Flower IAT. We anticipate that the effect of completing multiple IATs (vs. single IAT) will primarily be observed among participants who complete the Black–White IAT last (vs. first). Thus, in Experiment 3, we manipulate both the number (i.e., number of IATs) and order of the IATs (type of IAT administered last).

Experiment 3 employed a 2 (Number of IATs; Single vs. Multiple) × 2 (Type of IAT Administered Last; Black–White IAT vs. Insect–Flower IAT) between-subjects design. Participants were first randomly assigned to complete one of two IATs—Black–White IAT or Insect–Flower IAT—and were then either assigned to complete the second, alternative IAT or did not complete a second IAT. Thus, Experiment 3 included four conditions: (a) B/W test only, (b) I/F test only, (c) B/W followed by I/F test, and (d) I/F followed by B/W test.

Participants

We recruited 1,162 White U.S. citizens as volunteers from the Project Implicit educational website (https://implicit.harvard.edu/implicit/) to participate in Experiment 3. Of these, 144 were excluded from analyses for failing to complete the measures following completion of the first IAT to which they were assigned. A total of 1,018 White U.S. citizens completed the full survey ($M_{age} = 38.36$, $SD = 16.95$). Most participants have earned at least a bachelor's degree (62.2%) and self-identified as female (69%; male = 31%). The sample also skewed liberal in ideological self-placement ($M = 4.80$, $SD = 1.76$ on a 7-point scale ranging from 1 = *strongly conservative* to 7 = *strongly liberal*). We conducted a power analysis for small and medium-sized effects based on this sample of White U.S. citizens.

For a significant two-way interaction between the number of IATs and type of IAT administered last, our estimate is that this experiment had at least 35% power to detect a Cohen's f of 0.10% and 98% power to detect a Cohen's f of 0.25. To detect mean-level differences between the single and multiple IAT condition for people who first completed the Black–White IAT ($n = 526$) or the Flower–Insect IAT ($n = 492$), separately, our estimate is that this experiment had at least 59% power to detect a Cohen's d of 0.2% and 99% power to a Cohen's d of 0.5 or higher within each IAT condition. Finally, our sensitivity power analysis for the mediation analysis estimates that we had at least 88% power to detect an indirect effect of Cohen's d of 0.2 or higher when comparing I/F then B/W Test and B/W test only conditions, collapsed across the IAT condition ($N = 485$).

Transparency and Openness

All data exclusions (if any), all manipulations, and all measures in Experiment 3 are fully described. Data were analyzed using STATA, Version 17.0. Ethical approval was granted by the authors' Institutional Ethics Committee for all research activities reported in Experiment 3. The data and data syntax for the analysis are available at https://osf.io/7tfcs/?view_only=0819a809d2964aec8fe44ff998db601c.

Measures and Procedure

Experiment 3 employs the same procedure as Experiment 2, except for the inclusion of the two new conditions in which participants complete the Insect–Flower IAT alone or after completing the Black–White IAT. Furthermore, in Experiment 2, for participants who completed the Insect–Flower IAT prior to the Black–White IAT, we also manipulated whether or not participants were provided with information that explicitly connects the logic underlying the methodology of both tests and which emphasizes the role of cultural learning in the acquisition of implicit attitudes. We did not observe significant differences among participants who did or did not receive these instructions in Experiment 2, although future research should unpack the features of these instructions to better understand what might be driving these effects. For the purposes of the current research, what matters most is that the effect size was larger when information that explicitly connects the logic underlying the methodology of both tests is provided. For this reason, all participants who were administered multiple IATs were provided information connecting the underlying logic of both IATs, after completing the dependent measures following the first IAT, but before beginning the second IAT. We used the same measures in Experiment 3 as for Experiment 2, which were administered after each IAT, following feedback. All other measures, manipulations, and exclusions are otherwise fully reported.

Results

Table 3 provides the means, standard deviations, alphas, and intercorrelations of all measures included in this analysis. The overall means reported in Table 3 indicate that participants showed a pro-Flower or pro-White implicit association in respective IAT conditions. The direction and magnitude of preference observed in each test are consistent with results observed in previous research and in Experiments 1 and 2 (and are significantly different from zero, $ps < .001$). We again observe that the Black–White IAT results are not significantly correlated with bias awareness (Black–White D, $r = .02, p = .57$). As in Experiments 2 and 3, including IAT *D* scores as a covariate did not statistically change our conclusions. However, unlike Experiments 1 and 2, performance on the Insect–Flower IAT was associated with higher levels of bias awareness (Insect–Flower D, $r = .17$, $p < .001$), suggesting that those who report more pro-Flower implicit preferences were also more likely to report higher levels of Bias Awareness. Importantly, increased levels of bias awareness were also positively and significantly correlated with more favorable assessment IAT-VU for both IATs and increased test score feedback acceptance (Black–White IAT assessment of IAT-VU, $r = .41$, $p < .001$, Black–White IAT feedback acceptance, $r = .22$, $p < .001$; Insect–Flower IAT assessment of IAT-VU, $r = .28, p < .001$, Insect–Flower IAT feedback acceptance, $r = .19, p < .001$). Additionally, increased levels of bias awareness correlated with increased support for actions against bias ($r = .29$, $p < .001$). That is, more favorable assessments of the IAT and increased acceptance of its results, not performance on the Black–White IAT itself, predict increased levels of Bias Awareness, which is significantly associated with increased Support for Actions Against Bias.

To examine the effects of our independent variable in Experiment 3, we conducted a two-way between-subjects ANOVA and, following significant interactions, examined the main effect of number of IATs separately for participants who completed the Black–White IAT or Insect–Flower IAT last.

Did Assessment IAT-VU (After) and Test Score Feedback Acceptance Vary as a Function of Number of IATs and Type of IAT Administered Last?

First, we examined the effect of our two independent variables (number of IATs, type of IAT administered last) on the two critical

Table 3
Mean, Standard Deviation, and Intercorrelation of Measures Assessed in Experiment 3

Variables	*M*	*SD*	Alpha	Bug IAT D	Race IAT D	Bug IAT-VU	Race IAT-VU	Race FB acceptance	Bug FB acceptance	Act against bias	Bias aware
Insect–Flower IAT D	0.60	0.39	—	—							
Black–White IAT D	0.36	0.43	—	−.01	—						
Race IAT-VU	5.08	1.35	.89	.18**	−.07	—					
Bug IAT-VU	5.22	1.22	.87	.16**	−.11**	.60**	—				
Race FB acceptance	4.46	1.79	—	.05	−.22**	.27**	.61**	—			
Bug FB acceptance	5.15	1.79	—	.24**	−.07	.55**	.35**	.19**	—		
Act against bias	5.98	1.06	.73	.08	−.13**	.31**	.34**	.10*	.14**	—	
Bias awareness	4.43	1.52	.87	.17**	.02	.28**	.41**	.22**	.19**	.29**	—

Note. All explicit measures are assessed on a 1–7 scale. Higher values correspond with pro-Flower/anti-Insect, pro-White/anti-Black, more favorable attitudes of test validity and utility, increased willingness to support action against bias, and increased acceptance of bias feedback and awareness of bias. The information contained in these tables is based on all available observations. IAT D = results of the Implicit Association Test; IAT-VU = assessment of test validity and utility, for each IAT; FB acceptance = acceptance of the results of the IAT; bias awareness = awareness of and concern about one's own implicit racial bias; act against bias = willingness to support actions that combat the consequences of bias; IAT = Implicit Association Test; VU = validity and utility.
* $p < .05$. ** $p < .01$.

dependent variables (assessment IAT-VU, test score feedback acceptance). Figure 5 graphically represents the mean of each dependent variable across conditions.

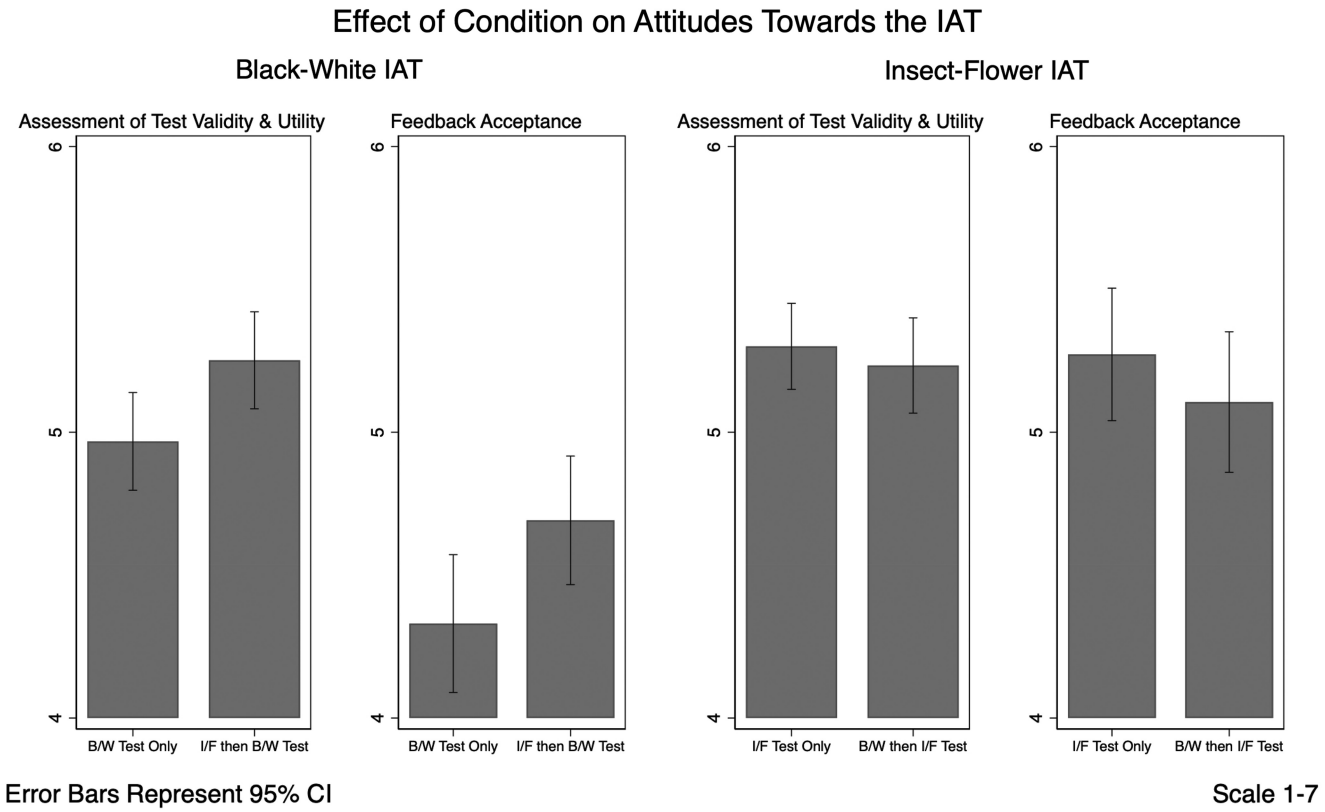
Assessment IAT-VU (After)

We observe a significant interaction between number of IATs and Type of IAT administered last, $F(3, 936) = 4.34$, $p = .038$, suggesting that the effect of multiple (vs. single) IATs may differ depending on the ordering of the IATs. We did not observe a significant effect of number of IATs for participants who completed the Insect–Flower IAT last, $t(450) = 0.28$, 95% CI for $M_{\text{difference}}$ [−0.16, 0.29], $p = .559$. However, for participants who completed the Black–White last (i.e., after the Insect–Flower IAT), we observed a significant effect of number of IATs, $t(486) = -2.32$, 95% CI for $M_{\text{difference}}$ [−0.53, −0.04], $p = .021$, Cohen's $d = 0.21$. Thus, participants who completed an Insect–Flower IAT before a Black–White IAT reported significantly more favorable assessments of the Black–White IAT than participants who only completed the Black–White IAT. We did not observe a differential effect on Assessment IAT-VU for people who completed the Black–White IAT first and Insect–Flower IAT second, compared to those who only completed the Insect–Flower IAT.

Test Score Feedback Acceptance

We observe a significant interaction between number of IATs and type of IAT administered last, $F(3, 874) = 4.82$, $p = .028$. This finding again indicates that the effect of multiple (vs. single) IATs is conditioned on the ordering of the IATs. We did not observe a significant effect of number of IATs for participants who completed the Insect–Flower IAT last, $t(421) = 097$, 95% CI for $M_{\text{difference}}$ [−0.17, 0.51], $p = .332$. However, for participants who completed the Black–White last (i.e., after the Insect–Flower IAT), we observed a significant effect of number of IATs, $t(453) = -2.15$, 95% CI for $M_{\text{difference}}$ [−0.69, −0.03], $p = .032$, Cohen's $d = 0.20$. Thus, participants who completed an Insect–Flower IAT before a Black–White IAT were more likely to accept the results of the Black–White IAT than participants who only completed the Black–White IAT. We did not observe a differential effect on test score feedback acceptance for people who completed the Black–White IAT first and Insect–Flower IAT second, compared to those who only completed the Insect–Flower IAT.

Figure 5
Between-Subject Comparison of Mean-Level Responses × Condition, Experiment 3



Note. B/W test only = participants only completed the Black–White IAT; I/F then B/W test = participants first completed the Insect–Flower IAT before the Black–White IAT; I/F test only = participants only completed the Insect–Flower IAT; B/W then I/F test = participants first completed the Insect–Flower IAT before the Black–White IAT. Higher values = more favorable attitudes and increased willingness to accept feedback. IAT = Implicit Association Test; B/W = Black/White; I/F = Insect/Flower.

Are There Indirect Effects on Bias Awareness and Support for Actions to Mitigate Bias?

Finally, we examined the indirect effects of multiple (vs. single) IATs on bias awareness and support for actions to mitigate bias, through increases in favorable assessment IAT-VU and test score feedback acceptance. To test this, we adopted the same procedure for estimating indirect effects as in Experiments 1 and 2. Because we only observed differences in our dependent variables as a function of the Number of IATs for participants who completed the Black–White (vs. Insect–Flower) IAT last, our analysis of indirect effects is limited to that condition. That is, we examine the indirect effects of I/F followed by B/W test (i.e., completed the Insect–Flower IAT first, followed by a Black–White IAT), compared to B/W test only (i.e., only completed a Black–White IAT).

Bias Awareness

With condition submitted as an independent variable (0 = *I/F then B/W test*, 1 = *B/W test only*) and assessment IAT-VU submitted as a mediator, the indirect effect of I/F then B/W test (vs. B/W test only) condition obtained significance on bias awareness ($b = -0.15$, $SE = 0.06$, 95% CI [−0.27, −0.03], $p = .016$). The direct effect of I/F then B/W test (vs. B/W test only) condition was not significant ($b = 0.03$, $SE = 0.13$, 95% CI [−0.22, 0.38], $p = .82$), nor was the total effect ($b = -0.12$, $SE = 0.14$, 95% CI [−0.39, 0.16], $p = .403$). That is, by increasing favorable Assessment IAT-VU, the I/F then B/W test (vs. B/W test only) condition indirectly increased bias awareness. Similarly, with test score feedback acceptance submitted as a mediator, the indirect effect of I/F then B/W test (vs. B/W test only) condition on bias awareness approached but did not reach statistical significance ($b = -0.06$, $SE = 0.04$, 95% CI [−0.14, 0.001], $p = .053$). However, the direct effect of I/F then B/W test (vs. B/W test only) was not significant ($b = -0.07$, $SE = 0.14$, 95% CI [−0.34, 0.21], $p = .646$), nor was the total effect ($b = -0.13$, $SE = 0.14$, 95% CI [−0.41, 0.15], $p = .36$). This pattern of results does not provide strong evidence that the I/F then B/W test (vs. B/W test only) indirectly increased bias awareness by increasing test score feedback acceptance; however, the effect of the experimental manipulation on bias awareness is fully mediated by its effects on assessments of the IAT.

Support for Actions to Mitigate Consequences of Bias

With condition submitted as an independent variable (0 = *I/F then B/W test*, 1 = *B/W test only*) and Assessment IAT-VU submitted as a mediator, the indirect effect of I/F then B/W Test (vs. B/W test only) condition obtained significance on Support for actions to mitigate the consequences of bias ($b = -0.08$, $SE = 0.04$, 95% CI [−0.16, −0.01], $p = .025$). The direct effect of I/F then B/W test (vs. B/W test only) condition was not significant ($b = 0.06$, $SE = 0.09$, 95% CI [−0.12, 0.23], $p = .519$), nor was the total effect ($b = -0.03$, $SE = 0.10$, 95% CI [−0.21, 0.16], $p = .788$). That is, by increasing favorable assessment IAT-VU, the I/F then B/W test (vs. B/W test only) condition indirectly increased support for actions to combat the consequences of bias. With test score feedback acceptance submitted as a mediator, the indirect effect of I/F then B/W test (vs. B/W test only) condition on support for actions to mitigate the consequences of bias approached but did not reach statistical significance ($b = -0.04$, $SE = 0.02$, 95% CI [−0.07, 0.002], $p = .067$). However, the direct effect of I/F then B/W test (vs. B/W test only) was not significant ($b = 0.01$, $SE = 0.10$, 95% CI [−0.19, 0.19], $p = .962$), nor was the total effect ($b = -0.03$, $SE = 0.10$, 95% CI [−0.22, 0.16], $p = .761$). These findings provide weak support for the expectation that I/F then B/W test (vs. B/W test only) would indirectly increase support for actions to mitigate the consequences of bias by increasing test score feedback acceptance; that is, the effect of the experimental manipulation on support for actions to combat bias is fully mediated by its effects on assessments of the IAT.

Discussion

Evidence of automatic bias was first demonstrated in experimental psychology through empirical demonstrations by Devine (1989) and Dovidio et al. (1986). The concepts of "implicit stereotypes" (Banaji & Prentice, 1994) and "implicit bias" (Greenwald & Banaji, 1995) were introduced into the scientific literature to capture growing evidence of a disparity between conscious and less conscious attitudes and beliefs. Since 1998, the IAT has been publicly available via the Internet (first at Yale University and later at Harvard University (https://www.implicit.harvard.edu), with usage on an unexpected and unprecedented scale; over 4.5 million tests of race bias alone have been analyzed (Charlesworth & Banaji, 2022; Morehouse & Banaji, 2024). In the 1990s, it was inconceivable that the term "implicit bias" would so fully permeate public consciousness and become a neologism. Now, every day, there are calls for implicit bias education in the popular press and social media (see Kella et al., 2023). Yet, very little is known about how to deliver high-quality education about implicit bias (see Banaji & Dobbin, 2023).

Learning and Educating About Bias

To educate the public effectively about concepts that were first discovered and introduced by experimental psychology, it is critical to identify strategies and methods that can aid accurate public understanding. For example, a growing number of papers have demonstrated that regional levels of IAT racial bias predict race differences in several socially significant outcomes, including lethal use of force by police, health outcomes, school discipline, and economic measures (see Charlesworth & Banaji, 2022 for a summary). Each paper reports that the greater the implicit race bias in a region (e.g., a county or state), the worse the outcomes for Black Americans relative to White Americans. Studies like these necessitate that implicit bias, however hidden, must be understood and acknowledged if society is to move toward a true understanding of itself and its stated commitment to equality and justice for all citizens.

Learning that one may harbor bias toward members of stigmatized groups is a necessary (although insufficient) step toward addressing inequity. However, evidence of bias is easy to reject because feedback about one's bias can be experienced as an affront to one's moral integrity, rationality, and competence (e.g., Crandall et al., 2002; Dovidio & Gaertner, 2000; Frantz et al., 2004). Without compassionate and clear communication of what implicit bias is, especially its ordinary and pervasive nature, even well-intended antibias programs geared toward raising awareness can create resistance, make people distrustful of the intervention, the science that underlies it, and the organization that promotes it (Banaji & Dobbin, 2023; Moskowitz & Vitriol, 2021). Psychological

dynamics that motivate rejection of the science of implicit bias also frustrate learning. Clearly, education and training programs that backfire can undermine the success with which individuals and organizations are willing and able to align their behavior or policies with their purported egalitarian values and goals (Dobbin & Kalev, 2016).

The aim of this program of work was to test a strategy for reducing resistance to evidence of implicit bias, in order to increase awareness of bias and support for action to combat its consequences for behavior and policies. Throughout the past 20 years, millions of volunteers to the Project Implicit website have taken one of several IATs. At the site, many visitors learn that they hold biased cognitive associations that may be hidden from conscious awareness or control. In this context, across three experiments using five independent samples, we demonstrated that leveraging a desire for consistency prior to receiving feedback can promote a more accurate understanding of oneself. Specifically, we elicited initial commitment to a favorable view of a test of implicit attitudes, and these simple prefeedback commitments operated as rational constraints and powerful psychological motivations that preempted the willingness to resist an unwelcome discovery about one's mind.

Like prior research on reactions to bias feedback (e.g., Howell et al., 2013), the present experiments demonstrated that participants were less inclined to accept evidence of racial bias relative to evidence about flora-fauna attitudes. Yet, after evaluating a test of implicit racial attitudes prior to feedback, participants in Experiment 1 adopted more favorable attitudes toward the very same test, postfeedback. We replicate these observations in Experiments 2 and 3 by having participants evaluate an Insect–Flower IAT before completing and receiving feedback from a Black–White IAT. In contrast, prefeedback commitment had no effect on test attitudes or acceptance of the results for the Insect–Flower IAT—a testing domain in which the feedback is unlikely to challenge importantly held values or social norms. Importantly, by increasing favorable attitudes toward the test and acceptance of test results (in Experiment 3), our intervention increased both awareness of bias and support for actions to mitigate bias; a first and necessary step toward prejudice regulation and policy reform (e.g., Bezrukova et al., 2016; Czopp et al., 2006; Lai et al., 2016; Perry et al., 2015; Vitriol, Banaji, & Lowe, 2024). This is no small matter, as a major goal of bias education and antibias training is to create this kind of awareness and subsequent motivation to act among those who consciously hold egalitarian goals (Moskowitz & Li, 2011).

Thus, a major contribution of our studies is its potential application to educational programs. Efforts to produce effective implicit bias education based on the best science available are underway (see as an example, outsmartingimplicitbias.org), with a cornerstone being the delivery and explanation of tests of implicit bias. One direct implication of our findings is that educational programs on implicit bias can benefit from leveraging the virtue of consistency, and our studies provide examples of subtle and effective ways to do so. By eliciting initial commitment to a favorable view of a test of implicit attitudes, prior to delivering unflattering results that may activate resistance, such an education program is likely to be accompanied by an increase in awareness of bias and support for action to combat its consequences for behavior and policies (see Vitriol, Banaji, & Lowe, 2024 as an example).

While our studies demonstrate that receptivity to feedback about bias can facilitate learning about and support for policies to combat bias, we did not examine the processes that underpin change in bias behavior or structural outcomes. Instead, the process examined by the current set of studies focused on the first step of a longer path toward an understanding of one's mind that can translate into behavior; how such gains lead to sustained change in behavior or structural outcomes remains a task for future research to address (see Carter et al., 2020). We expect that changing behavior and institutions over the long term will require additional work at multiple levels, sustained over protracted timeframes (Moskowitz & Vitriol, 2021). In our view, the foundation of these changes is likely to begin with increasing awareness of the cause and consequences of bias, which will depend on successfully managing resistance that arises when learning about bias. A benefit of the present experiments is that awareness was created without pedantic lectures or argumentation; it was created simply by allowing the individual to recognize that it is cognitively inconsistent to accept the outcome of Test 1 (Insect–Flower test) but not the outcome of a parallel Test 2 (Black–White test).

Activating Consistency Motivations to Navigate Resistance in Bias Education

What explains the effect of our intervention on increased acceptance of bias feedback? The overall pattern of results across experiments suggests that initial evaluations—whether of the less threatening Flowers–Insects IAT or prefeedback evaluations of the Black–White IAT—were more positive, and that these initial positive evaluations were then reflected in subsequent evaluations and attitudes postfeedback. These observations clearly align with major perspectives and findings in social psychology for the past half-century and more. The critical mechanism is a drive toward consistency (Bem, 1972; Festinger & Carlsmith, 1959; Stone et al., 1994).

Psychologists have long described the desire for consistency between beliefs, attitudes, identities, and behaviors as a basic human motivation (i.e., Festinger, 1957; Heider, 1959; for a review, see Gawronski, 2012). Festinger spoke of this motivation with his concept of cognitive-dissonance, arguing that because inconsistency is uncomfortable and normatively undesirable, humans seek to align their beliefs and behavior with each other (Festinger, 1957). Similarly, Heider (1959) proposed that humans desire balance in their cognitive evaluations of the self, others, and objects. More recently, Greenwald et al. (2003) demonstrated that this desire for balance can operate at the automatic, implicit level of cognition. Regardless of whether it is experienced due to an imbalanced cognitive system or conflicts between beliefs and behaviors, inconsistency is usually experienced as both distressing and aversive (Burke, 2006). The discomfort that arises from inconsistency is motivating—it elicits compensatory responses to reduce or resolve the dissonance or imbalance through a change in cognition and behavior, helping one to acquire a state of consonance and a more stable basis for effective action and goal pursuits (Banaji & Bhaskar, 2000; DeMarree et al., 2015; Gollwitzer & Bayer, 1999; Harmon-Jones et al., 2009). This compensatory motivation following inconsistency is critical to understanding the nature of resistance to learning about bias and, more importantly, the effects of our intervention.

In these studies, we subtly but effectively induced recognition of one's prior view on the validity of a measure. Once that view exists within conscious awareness, consistency theories predict that behavior should line up. And indeed, given the difference in outcomes between the experimental and control conditions, we see that cognitive consistency prevailed. Behavioral change in this context is difficult, not least because it is impossible to revise past responses that were offered in response to the simpler test (Insect–Flower) or prior to knowing one's performance. By creating the conditions that elicited prefeedback commitments to a favorable view of the test of implicit attitudes, we made salient a potential conflict with the desire to resist or ignore unwelcome evidence of bias. This disparity, between a favorable view of the test and the unflattering nature of its results, rendered participants less able or willing to employ strategies like denial, justification, or rationalization as a means of resolving the inconsistency between a desirable self-image and the threat posed by the feedback. As a result, increased acceptance of the results and sincere commitment to egalitarian goals became a more efficient pathway to reducing the discomfort of learning that one has acted with bias. The most likely mechanism underlying the obtained effect is that by making one's previous response memorable, we engaged the value that modern humans place on acting consistently–consistently with what one has previously espoused, how one has previously behaved, and in line with one's consciously accessible values.

A limitation of the present research is that we did not directly test the underlying mechanisms of consistency motivations in this context by, for example, measuring the experience of cognitive-dissonance (but see Sackett, 2021) or cognitive-imbalance (nor do we measure the type of aversive arousal or impression-management concerns that are central to the motivations described by theories of consistency). Future research should more directly investigate these underlying processes to better understand why and how the effects of our experiments occur, as well as its downstream effects on spontaneously generated bias awareness and action.

Constraints on Generalizability and Other Limitations

Among additional limitations of these studies is, of course, the concern about the generalizability of our observations (Vitriol et al., 2019). Because participants self-select into Project Implicit, whether our observations will replicate on more representative samples remains unknown. Project Implicit is used by many organizations that direct their entire membership to the site for an educational experience, so the demographic characteristics of individuals who are channeled into the research site are diverse. Data from other studies using Project Implicit participants compare well with data collected from a representative sample (Morehouse et al., 2022). However, it is possible that this method for increasing bias acceptance may not generalize to those individuals who do not explicitly endorse egalitarian values, nor to those who reject the idea that Black Americans face discrimination. Indeed, a commitment to equality (which may be over-represented in our sample) as a moderator of this effect should be examined in future research.

Further, it would be useful to test how the equality imperative, especially as it concerns the fate of Black Americans, emerges in groups other than White Americans. For example, Asian Americans show the same degree of implicit anti-Black bias as White Americans. Still, we know far less about their desire to appear egalitarian in the context of Black–White attitudes. The data of Black Americans are also of interest, as Black Americans express satisfaction when their test score shows pro-Black implicit attitudes (the opposite of White Americans in the sense that Black Americans are comfortable with showing in-group preference). If there is resistance to the test result in Black Americans, it is likely to be when their data reveal pro-White attitudes (Howell et al., 2015). These variations in racial or ethnic group differences in receiving feedback are intriguing and worthy of investigation in the context of who is persuaded to accept the validity of the test and accuracy of its results; that is, who is willing to accept evidence of bias.

Future research should also seek to develop and validate standardized measures of defensive reactions to implicit bias feedback. There are currently no validated measures in the published literature in the conventional sense, and the current set of studies is no exception. Nonetheless, the measures we used in our experiments have high levels of face validity and internal validity, and are highly similar to previous implementations in the research literature. For example, conceptually similar items are used in all of the reported studies in Vitriol and Moskowitz (2021), Rothman et al. (2022), Howell and Ratliff (2016), Howell et al. (2015, 2017). Furthermore, one indication that our measure captures the constructs of interest is in the mean-level variability across IATs. We consistently observe that perceptions of test validity and utility and test feedback acceptance are higher in the Flower–Insect IAT compared to the Black–White IAT. Additional evidence that our measures of defensive reactions capture the constructs of interest is that, as we would expect, they predict awareness of bias and support to combat the consequences of bias. This consistent pattern of responses across multiple samples and experimental paradigms suggests that our measures are appropriate for the hypotheses examined in these studies. Still, a scientific understanding of this phenomenon can be advanced with more rigorous and programmatic development of validated measures.

Conclusion

Together, our findings demonstrate that prefeedback evaluation can attenuate the motivation to devalue unflattering feedback from a test of implicit attitudes in socially sensitive domains, thereby increasing bias awareness and support for actions to combat bias; critical outcomes for promoting learning about diversity. By activating motivations to maintain consistent beliefs toward evidence of bias, bias education can promote more rational and moral judgment in the domain of intergroup prejudice—a context in which self-rationalization and justification have all too often been the common resolution to a dilemma that has bedeviled American society for generations.

References

Adams, V. H., Devos, T., Rivera, L. M., Smith, H., & Vega, L. A. (2014). Teaching about implicit prejudices and stereotypes: A pedagogical demonstration. *Teaching of Psychology*, *41*(3), 204–212. https://doi.org/10.1177/0098628314537969

Aneja, A., Luca, M., & Reshef, O. (2023). *The benefits of revealing race: Evidence from minority-owned local business*. Harvard Business School, Working Paper 23-042.

Aronson, E. (1999). Dissonance, hypocrisy, and the self-concept. In E. Harmon-Jones & J. Mills (Eds.), *Cognitive dissonance: Progress on a pivotal theory in social psychology* (pp. 103–126). American Psychological Association. https://doi.org/10.1037/10318-005

Axt, J. R., Casola, G. M., & 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

Banaji, M. R., & Bhaskar, R. (2000). Implicit stereotypes and memory: The bounded rationality of social beliefs. In D. L. Schacter & E. Scarry (Eds.), *Memory, brain, and belief* (pp. 139–175). Harvard University Press.

Banaji, M. R., & Dobbin, F. (2023) *Why DEI training doesn't work—And how to fix it—Division of social science*. https://www.wsj.com/business/c-suite/dei-training-hr-business-acd23e8b?st=5e2l3cu8vu431pc&reflink=desktopwebshare_permalink

Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Delacorte Press.

Banaji, M. R., Hardin, C., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, *65*(2), 272–281. https://doi.org/10.1037/0022-3514.65.2.272

Banaji, M. R., & Prentice, D. A. (1994). The self in social contexts. *Annual Review of Psychology*, *45*(1), 297–332. https://doi.org/10.1146/annurev.ps.45.020194.001501

Bem, D. J. (1972). Self-perception theory11development of self-perception theory was supported primarily by a grant from the National Science Foundation (GS 1452) awarded to the author during his tenure at Carnegie-Mellon University. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). Academic Press. https://doi.org/10.1016/S0065-2601(08)60024-6

Bergsieker, H. B., Shelton, J. N., & Richeson, J. A. (2010). To be liked versus respected: Divergent goals in interracial interactions. *Journal of Personality and Social Psychology*, *99*(2), 248–264. https://doi.org/10.1037/a0018474

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. https://doi.org/10.1037/bul0000067

Bobo, L. (2001). Racial attitudes and relations at the close of the twentieth century. In *America becoming: Racial trends and their consequences* (Vol. 1, pp. 264–301). National Academies Press. https://people.umass.edu/cnle/soc361/docs/ab1-9.pdf

Burke, P. J. (2006). Identity change. *Social Psychology Quarterly*, *69*(1), 81–96. https://doi.org/10.1177/019027250606900106

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

Carnes, M., Devine, P. G., Baier Manwell, L., Byars-Winston, A., Fine, E., Ford, C. E., Forscher, P., Isaac, C., Kaatz, A., Magua, W., Palta, M., & Sheridan, J. (2015). The effect of an intervention to break the gender bias habit for faculty at one institution: A cluster randomized, controlled trial. *Academic Medicine*, *90*(2), 221–230. https://doi.org/10.1097/ACM.0000000000000552

Carter, E. R., Onyeador, I. N., & Lewis, N. A. (2020). Developing & delivering effective anti-bias training: Challenges & recommendations. *Behavioral Science & Policy*, *6*(1), 57–70. https://doi.org/10.1177/237946152000600106

Charlesworth, T. E. S., & Banaji, M. R. (2022). Patterns of implicit and explicit attitudes: IV. Change and stability from 2007 to 2020. *Psychological Science*, *33*(9), 1347–1371. https://doi.org/10.1177/09567976221084257

Cialdini, R. (2008). *Influence: Science and practice* (5th ed.). Allyn and Bacon.

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. https://doi.org/10.1037/0022-3514.82.3.359

Czopp, A. M., Monteith, M., & Mark, A. Y. (2006). Standing up for 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

DeMarree, K. G., Briñol, P., & Petty, R. E. (2015). Reducing subjective ambivalence by creating doubt: A metacognitive approach. *Social Psychological and Personality Science*, *6*(7), 731–739. https://doi.org/10.1177/1948550615581497

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*(6), 1267–1278. https://doi.org/10.1016/j.jesp.2012.06.003

Dobbin, F., & Kalev, A. (2016). Why diversity programs fail. *Harvard Business Review*, *94*(7). https://hbr.org/2016/07/why-diversity-programs-fail

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, *22*(1), 22–37. https://doi.org/10.1016/0022-1031(86)90039-9

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*(4), 315–319. https://doi.org/10.1111/1467-9280.00262

Duguid, M. M., & Thomas-Hunt, M. C. (2015). Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes. *Journal of Applied Psychology*, *100*(2), 343–359. https://doi.org/10.1037/a0037908

Eisenhauer, J. G. (2021). Meta-analysis and mega-analysis: A simple introduction. *Teaching Statistics*, *43*(1), 21–27. https://doi.org/10.1111/test.12242

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, *58*(2), 203–210. https://doi.org/10.1037/h0041593

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T. L., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology*, *72*, 133–146. https://doi.org/10.1016/j.jesp.2017.04.009

Frantz, C. M., Cuddy, A. J. C., Burnett, M., Ray, A., & Hart, A. (2004). A threat in the computer: The race Implicit Association Test as a stereotype threat experience. *Personality and Social Psychology Bulletin*, *30*(12), 1611–1624. https://doi.org/10.1177/0146167204266650

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition*, *30*(6), 652–668. https://doi.org/10.1521/soco.2012.30.6.652

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*, *14*(4), 574–595. https://doi.org/10.1177/1745691619826015

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549. https://doi.org/10.1111/spc3.v10.10

Gollwitzer, P. M., & Bayer, U. (1999). Deliberative versus implemental mindsets in the control of action. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 403–422). The Guilford Press.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist*, *35*(7), 603–618. https://doi.org/10.1037/0003-066X.35.7.603

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

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

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Harmon-Jones, E., Amodio, D. M., & Harmon-Jones, C. (2009). Action-based model of dissonance: A review, integration, and expansion of conceptions of cognitive conflict. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 41, pp. 119–166). Elsevier Academic Press.

Heider, F. (1959). *The psychology of interpersonal relations*. Wiley. https://doi.org/10.2307/2572978

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., Collisson, B. D., Crysel, L., Garrido, C. O., Newell, S. M., Cottrell, C. A., Smith, C. T. S., & Shepperd, J. A. (2013). Managing the threat of impending implicit attitude feedback. *Social Psychological and Personality Science*, *4*(6), 714–720. https://doi.org/10.1177/1948550613479803

Howell, J. L., Gaither, S. E., & Ratliff, K. A. (2015). Caught in the middle: Defensive responses to IAT feedback among Whites, Blacks, and Biracial Black/Whites. *Social Psychological and Personality Science*, *6*(4), 373–381. https://doi.org/10.1177/1948550614561127

Howell, J. L., & Ratliff, K. A. (2016). 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.1216

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

Hughes, S., De Houwer, J., Mattavelli, S., & Hussey, I. (2020). The shared features principle: If two objects share a feature, people assume those objects also share other features. *Journal of Experimental Psychology: General*, *149*(12), 2264–2288. https://doi.org/10.1037/xge0000777

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, *25*(6), 881–919. https://doi.org/10.1111/j.1467-9221.2004.00402.x

Kella, S., Morehouse, K., & Banaji, M. R. (2023). *Implicit bias in the public eye: Using Google alerts to determine public sentiment* [Poster presentation]. Annual Meetings for Society for Personality and Social Psychology, Atlanta, Georgia, United States.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36*(2), 249–277. https://doi.org/10.1207/S15327906MBR3602_06

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

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science*, *22*(12), 1472–1477. https://doi.org/10.1177/0956797611427918

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., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: On the development and operation of cues for control. *Journal of Personality and Social Psychology*, *83*(5), 1029–1050. https://doi.org/10.1037/0022-3514.83.5.1029

Monteith, M. J., Devine, P. G., & Zuwerink, J. R. (1993). Self-directed versus other- directed affect as a consequence of prejudice-related discrepancies. *Journal of Personality and Social Psychology*, *64*(2), 198–210. https://doi.org/10.1037/0022-3514.64.2.198

Monteith, M. J., & Mark, A. Y. (2009). The self-regulation of prejudice. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 507–523). Psychology Press.

Morehouse, K. N., & Banaji, M. R. (2024). The science of implicit race bias: Evidence from the Implicit Association Test. *Daedalus*, *153*(1), 21–50. https://doi.org/10.1162/daed_a_02047

Morehouse, K. N., Kurdi, B., Hakim, E., & Banaji, M. R. (2022). When a stereotype dumbfounds: Probing the nature of the surgeon = male belief. *Current Research in Ecological and Social Psychology*, *3*, Article 100044. https://doi.org/10.1016/j.cresp.2022.100044

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*. Taylor & Francis.

Moskowitz, G. B. (2002). Preconscious effects of temporary goals on attention. *Journal of Experimental Social Psychology*, *38*(4), 397–404. https://doi.org/10.1016/S0022-1031(02)00001-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

Myrdal, G. (1944). *An American dilemma* (Vol. 1). Routledge.

O'Brien, L. T., Crandall, C. S., Horstman-Reser, A., Warner, R., Alsbrooks, A., & Blodorn, A. (2010). But I'm no bigot: How prejudiced White Americans maintain unprejudiced self-images. *Journal of Applied Social Psychology*, *40*(4), 917–946. https://doi.org/10.1111/j.1559-1816.2010.00604.x

Onyeador, I. N., Hudson, S. T. J., & Lewis, N. A. (2021). Moving beyond implicit bias training: Policy insights for increasing organizational diversity. *Policy Insights from the Behavioral and Brain Sciences*, *8*(1), 19–26. https://doi.org/10.1177/2372732220983840

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

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. https://doi.org/10.1037/0022-3514.75.3.811

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research*

Methods, Instruments & Computers, *36*(4), 717–731. https://doi.org/10.3758/BF03206553

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences*, *11*(1), 37–43. https://doi.org/10.1016/j.tics.2006.11.001

Regner, 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 Behavior*, *3*(11), 1171–1179. https://doi.org/10.1038/s41562-019-0686-3

Rothman, N. B., Vitriol, J. A., & Moskowitz, G. B. (2022). Internal conflict and prejudice-regulation: Emotional ambivalence buffers against defensive responding to implicit bias feedback. *PLoS ONE*, *17*(3), Article e0264535. https://doi.org/10.1371/journal.pone.0264535

Sackett, A. (2021). *Couldn't be me: Dissonance and defensive responding in bias interventions* [Master's thesis, Lehigh University].Sommers, S. R., & Norton, M. I. (2006). Lay theories about White racists: What constitutes racism (and what doesn't). *Group Processes & Intergroup Relations*, *9*(1), 117–138. https://doi.org/10.1177/1368430206059881

Stone, J. (2001). Behavioral discrepancies and the role of construal processes in cognitive dissonance. In G. B. Moskowitz (Ed.), *Cognitive social psychology: The Princeton symposium on the legacy and future of social cognition* (pp. 41–58). Lawrence Erlbaum Associates.

Stone, J., Aronson, E., Crain, A. L., Winslow, M., & Fried, C. (1994). Inducing hypocrisy as a means of encouraging young adults to use condoms. *Personality and Social Psychology Bulletin*, *20*(1), 116–128. https://doi.org/10.1177/0146167294201012

Uhlmann, E. L., & Nosek, B. A. (2012). My culture made me do it: Lay theories of responsibility for automatic prejudice. *Social Psychology*, *43*(2), 108–113. https://doi.org/10.1027/1864-9335/a000089

Vitriol, J. A. (2016). *The (In) Egalitarian self: On the motivated rejection of implicit racial bias* [Doctoral dissertation]. University of Minnesota.

Vitriol, J. A., Banaji, M. B., & Lowe, R. (2024). Change in attitudes and beliefs about implicit bias education: A demonstration among members of a police department. *Philosophical Psychology*. Advance online publication. https://doi.org/10.1080/09515089.2023.2296585

Vitriol, J. A., Larsen, E. G., & Ludeke, S. G. (2019). The generalizability of personality effects in politics. *European Journal of Personality*, *33*(6), 631–641. 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. (2024). Less bias, yet more defensive: The role of control processes. *Journal of Experimental Psychology: Applied*, *30*(1), 108–119. https://psycnet.apa.org/buy/2023-51925-001

Received February 15, 2023
Revision received February 16, 2024
Accepted March 13, 2024 ■