Reducing defensive responding to implicit bias feedback: On the role of perceived moral threat and efficacy to change

Article *in* Journal of Experimental Social Psychology · September 2021

DOI: 10.1016/j.jesp.2021.104165

CITATIONS
0

READS
15

2 authors:

Joseph A Vitriol
Stony Brook University
26 PUBLICATIONS **215** CITATIONS
SEE PROFILE

Gordon B. Moskowitz
Lehigh University
59 PUBLICATIONS **5,524** CITATIONS
SEE PROFILE

Journal of Experimental Social Psychology 96 (2021) 104165

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Reducing defensive responding to implicit bias feedback: On the role of perceived moral threat and efficacy to change☆

Joseph A. Vitriol [a,b,*], Gordon B. Moskowitz [c]

[a] Department of Psychology, Harvard University, United States of America
[b] Department of Political Science, Stony Brook University, United States of America
[c] Department of Psychology, Lehigh University, United States of America

ARTICLE INFO

Keywords:
Stereotyping
Prejudice
Implicit cognition
Attitudes
Self
Intergroup dynamics
Interventions

ABSTRACT

The last decade has seen a rush to address the causes and consequences of bias in applied contexts across the world. When and why might these initiatives promote attitudes and behavior that align with egalitarian goals? A common assumption is that increasing awareness of bias can motivate control over prejudiced responding. However, learning that one's actions are biased is threatening, and often motivates a range of self-protective responses to buffer that threat. In the current research, we tested a strategy for reducing such defensive responding and increasing the kind of awareness central to contemporary theories of prejudice regulation and egalitarian behavior. Four experiments ($N > 2500$) and a mini meta-analysis demonstrate that interventions that (a) decrease perceived moral blameworthiness for having bias and (b) increase the perceived ability to control bias, can reduce defensive responding and increase awareness both in the short-term and approximately 6 months later. Interventions that minimize threat and facilitate efficacy can motivate increased bias awareness and commitment to egalitarian values.

A substantial body of research across the social and behavioral sciences has revealed that individual-level biases in psychology can contribute to error in social judgment and help to maintain inequitable social relations (e.g., Kurdi et al., 2019). In response, the last decade has seen a rush to address the causes and consequences of bias in applied contexts across the world (e.g., Banaji & Greenwald, 2016; Hansen, 2003; Jost et al., 2009; Lai et al., 2014; Lipman, 2018). These efforts commonly involve interventions designed to make people aware of their biases and the aggregate impact of such bias (e.g., Carter, Onyeador, & Lewis Jr., 2021; Hillard, Ryan, & Gervais, 2013; Sekaquaptewa, Takahashi, Malley, Herzog, & Bliss, 2019; Stone, Moskowitz, Zestcott, & Wolsiefer, 2019). When and why might these initiatives promote change in attitudes and behavior that more closely align with egalitarian values and goals?

An essential element to the pursuit of any goal, egalitarian or otherwise, or the improvement of the self in any domain, is feedback. Feedback arrives both intentionally through the individual seeking it out (e.g., Carver & Scheier, 1998), and unintentionally through the (often unsolicited) remarks of others (e.g., Stone, 2001). The logic to many anti-bias interventions is that people often lack awareness of if, or how, their beliefs and actions are biased (e.g., Carter et al., 2021; Pronin, Lin, & Ross, 2002). Interventions are commonly designed to create awareness of bias, through feedback, that can promote the regulation of bias (e.g., Czopp, Monteith, & Mark, 2006; Moskowitz & Li, 2011; Perry, Murphy, & Dovidio, 2015). This is seemingly a reasonable expectation given three facts. First, many individuals already possess egalitarian goals (e.g., Moskowitz & Li, 2011). Second, the triggering of those goals and control over bias follows awareness (e.g., Axt & Casola, 2018; Forscher, Mitamura, Dix, Cox, & Devine, 2017; Monteith & Mark, 2009; Moskowitz, Gollwitzer, Wasel, & Schaal, 1999; Perry et al., 2015). Third, awareness can lead to bias reduction in real-world contexts (e.g., Carnes et al., 2015; Devine, Forscher, Austin, & Cox, 2012; Parker, Monteith, Moss-Racusin, & Van Camp, 2018; Regner, Thinus-Blanc, Netter, Schmader, & Huguet, 2019; Stone et al., 2019).

However, feedback not only motivates goal-pursuits, but can, at times, engender defensiveness that hinders goal-related performance, especially in domains that challenge one's integrity in the eyes of the self or others. Learning that one's cognitions, beliefs, or actions are biased is

☆ This paper has been recommended for acceptance by Dr Vanessa Bohns.

* Corresponding author at: Department of Political Science at SUNY, Stony Brook University, 100 Nicolls Road, Stony Brook, NY 11794, United States of America.
E-mail addresses: joevitriol@gmail.com, joseph.vitriol@stonybrook.edu (J.A. Vitriol).

https://doi.org/10.1016/j.jesp.2021.104165
Received 5 January 2021; Received in revised form 6 May 2021; Accepted 10 May 2021

inconsistent with personally held values and is socially stigmatizing (e.g., Crandall, Eshleman, & O'Brien, 2002; Devine, Monteith, Zuwerink, & Elliot, 1991). Many people are motivated to ignore, rationalize, or combat such threats to their egalitarian self-image (e.g., Dovidio & Gaertner, 2000; Frantz, Cuddy, Burnett, Ray, & Hart, 2004). Consequently, there are times that raising awareness via feedback in this domain has the opposite of the desired effect of regulating bias, triggering instead a constellation of self-protective responses comprising resentment, denial, denigrating the source of the feedback, and trivializing the importance of the domain in which the feedback occurs. Thus, anti-bias interventions that fail to account for this constellation of self-protective responses from raising awareness of bias may lead to the ironic and unfortunate consequence of worsening the situation rather than alleviating bias (for a review, see Moskowitz & Vitriol, 2021).

In the present set of experiments, we replicate existing findings demonstrating that people are defensive towards feedback about bias. We extend this work to examine how to mitigate this undesired effect of bias feedback. Our results reveal how to contextualize and frame feedback about bias to enable people to respond efficaciously in pursuit of their egalitarian goals, rather than defensively. Below, we report the results of a series of experiments demonstrating that defensive responding can be reduced by providing feedback about bias in a manner that (a) reduces perceived moral blameworthiness for having bias and (b) increases the person's perceived ability to control the expression of their bias. As a result of the decrease in defensiveness, increased bias awareness is observed, both in the short-term and approximately six months later. These findings suggest that the use of feedback to create bias awareness must take steps to minimize the backlash and defensiveness that an intervention can otherwise inadvertently promote. With the proper framing, feedback can successfully promote bias awareness, personal culpability, commitment to egalitarian goals (in intentions and actions).

1. When is feedback motivating versus a cause for defensiveness?

All theories of goal pursuit allow that feedback energizes responses that serve the goal. It stimulates action aimed at addressing the negative feedback (e.g., Bandura, 1991; Carver & Scheier, 1998; Higgins, Strauman & Klein, 1986; Wicklund & Gollwitzer, 1982). Negative feedback triggers an uncomfortable state, a tension, which lead to stronger intentions and goal commitment. When a low prejudice person receives feedback about bias, research shows that it can motivate control over bias and egalitarian goal pursuit. This is accomplished through responses such as: selective attention, inhibition of unwanted behaviors and stereotypic thoughts, heightened commitment to egalitarianism, increased accessibility of egalitarian goals, and conflict monitoring processes (e.g., Amodio, Devine, & Harmon-Jones, 2008; Czopp et al., 2006; Devine et al., 1991; Monteith, 1993; Monteith, Ashburn-Nardo, Voils, & Czopp, 2002; Moskowitz, 2002; Moskowitz et al., 1999; Moskowitz & Li, 2011).

However, there are times when feedback leads to disengagement from goals. Negative feedback is especially demotivating when it reduces one's perceived efficacy at goal pursuit, increases awareness of structural barriers to achieving the goal, or increases the value of alternative goals to be pursued instead (e.g., Klinger, 1975; Wrosch, Scheier, Carver & Schulz, 2003). Feedback indicating moral failure can sometimes lead to social anxiety and arousal for fear of appearing undesirable in the eyes of the self or others (e.g., Plant & Devine, 2003; Schlenker & Leary, 1982). This can make one avoidant of opportunities to pursue a goal, rather than vigilant to embrace such opportunities. These conditions under which feedback demotivates are often in place when the feedback is about one's stereotyping and prejudice.

1.1. Reduced efficacy and bias feedback

When feedback about bias indicates that one's bias is implicit, this suggests to the feedback recipient that such bias is difficult to control. Regardless of whether the feedback comes through a workshop, an intervention, or an online test (such as the Implicit Association Test, IAT), the characterization of bias as implicit may imply a reduced level of efficacy for prejudice-regulation. Research on automatic processing (e.g., Bargh, 2017) describes lack of control as one of its defining features. It is more difficult to control that which we cannot see and that which is habitual than that which is evident and deliberate. Indeed, the logic of increasing awareness of bias via feedback is to address the fact that it is difficult to control that which is unseen and unknown. Shining a light is expected to increase the desire to take action. Thus, shining a light on implicit bias through feedback can make it seen, but doing so may also communicate the difficulty in exerting control. This can be demotivating rather than motivating.

1.2. Normative/structural barriers and bias feedback

Similarly, implicit bias feedback suggests to its recipient that bias is a natural and common component of human cognition. As Daumeyer, Onyeador, Brown, and Richeson (2019) warn, it is easy for individuals to interpret discussions of "implicit bias as a common element of human cognition" as a call to trivialize personal culpability. The perception of implicit bias as widespread and common implies structural barriers to overcoming it. It implies that individual action will be irrelevant or trivial in the face of the shared biases unchecked in the entire community. While intended as a call for personal culpability, implicit bias feedback can instead reduce such culpability and lead to the perception that there are society-wide obstacles that cannot be scaled. If this is what implicit bias workshops are (unintentionally) communicating, they may undermine prejudice regulation by reducing feelings of culpability and efficacy in the individual.

1.3. Social anxiety and bias feedback

Feedback suggesting that one may be prejudiced can lead members of the majority group to avoid interactions with outgroup members (e.g., Dovidio & Gaertner, 2000; Frantz et al., 2004) and to pursue *other* goals that will compensate for the negative feedback, in domains where bias is irrelevant (they may seek self-affirmation; e.g., Steele, 1988). Few social labels are more aversive than the label of "racist" (Crandall et al., 2002), and being identified as such can cause shame or stigma. For example, Czopp et al. (2006) argued that when feedback comes as a confrontation it can be seen as accusatory and impugning. This is especially true if the feedback comes from a member of a minority group (e.g., Sidanius & Pratto, 2001). The anxiety and shame of feedback in this domain makes seeking affirmation in some other goal domain and avoiding the anxiety of addressing one's failed egalitarianism logical.[1]

Importantly, seeking affirmation in alternative goal domains may be insufficient to offset the anxiety and threat introduced by feedback about bias. In such instances, one may engage defensive responses, not merely avoidance. Defensiveness often takes the form of derogating the source of the feedback and questioning its credibility and objectivity (e.g., Shepperd, Malone, & Sweeny, 2008; Spencer, Fein, Wolfe, Fong, &

[1] Avoiding one's bias may yield other, unintended, consequences. While it may focus one on other goals that are less threatening, avoiding prejudice can require the monitoring of thoughts, feelings, and behavior (e.g., Richeson & Shelton, 2007; Vorauer & Kumhyr, 2001; Wegner, 1994). The effort required for monitoring prejudiced-responses increases cognitive demands, with negative consequences for subsequent goal pursuit and interpersonal functioning (e.g., Amodio & Hamilton, 2012; Kenrick, Sinclair, Richeson, Versoksy, & Lun, 2016; Richeson & Trawalter, 2008; Wegner, 1994).

Duinn, 1998). For example, shaming can turn into anger, making a person motivated to attack the impugning person and reject feedback (e.g., Baumeister & Campbell, 1999; Tangney, 1995). In a relevant line of work, Howell and colleagues assert that people who see themselves as low in prejudice are avoidant of feedback they *expect* to indicate implicit racial bias (Howell et al., 2013). When directly confronted with such feedback (thus, they are unable to avoid it), people were distressed (Howell, Gaither, & Ratliff, 2015). They responded to the feedback by impugning its credibility and accuracy. This was particularly true from people who underestimate their own levels of bias relative to the "average" American (Howell & Ratliff, 2016). Thus, people who may benefit the most from bias feedback may be particularly avoidant and defensive about bias feedback that challenges a socially desirable self-image. Rather, what results is defensiveness and the motivated rejection of implicit racial bias feedback (e.g., Howell, Redford, Pogge, & Ratliff, 2017).

2. Addressing the conditions that cause bias feedback to be demotivating

Feedback about bias is presumed to play an important role in producing long-term change in attitudes and behavior relating to egalitarianism (a decrease in bias). This increased egalitarianism is hypothesized to occur through feedback triggering 1) an increased awareness of bias, 2) personal culpability for responding to the bias, 3) commitment to the goal of egalitarianism and an intention to be more egalitarian, and 4) increased egalitarian action (e.g., Forscher et al., 2019; Howell et al., 2015; Lai et al., 2016; Moskowitz & Vitriol, in press). Feedback will not have this chain of desired consequences if it creates threat and activates defensiveness. How can we provide feedback about one's bias in a manner that minimizes defensiveness and that addresses the conditions that demotivate one's response to the feedback? Addressing these issues should allow the feedback to motivate egalitarian responding, as intended. Our approach to minimizing the demotivating effects of bias feedback benefitted from insights from the theory of planned behavior (Ajzcn, 1991).

2.1. Bias feedback that does not reduce efficacy

As Ajzcn (1991) reviews, goal pursuits are enhanced by removing the threat posed by feeling that one lacks the skills to take effective action (fear of being *unable* to overcome the shortcoming). Stone et al. (2019) report that a common response of intervention attendees is that they desire to not merely be made aware that they have bias—they also want to learn how to overcome it. Without being armed with tools to address or control bias, feedback about – and heightened awareness of – bias is too threatening. Indeed, preliminary research in our lab shows that people who naturally feel efficacious at controlling bias (people with ability to self-regulate) are less defensive to the bias feedback they receive (Vitriol, Calanchini, & O'Shea, 2020).

Using this logic, feedback about bias must be constructed in a way to communicate explicitly that, although the bias is implicit, it is controllable. Increasing perceived efficacy to control bias is critical to reducing defensiveness. The feedback must focus equally on shining a light on the existence of bias and creating personal *efficacy* at overcoming bias (Bandura, 1989). Efficacy can be achieved by training; Participants can be taught ways to inhibit or control bias (e.g., Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000; Mendoza, Gollwitzer, & Amodio, 2010; Stone et al., 2019). It can also be aided by creating in participants a general sense of having the ability to overcome bias. This can be done by an intervention that affirms the belief that implicit bias can be controlled via adaptive responding and behavioral change (e.g., Witte & Allen, 2000).

2.2. Bias feedback that does not imply normative/structural barriers

A second insight from Ajzcn (1991) is that norms dictate the individual's motivation to conform or dissent, with action most likely to be taken when contextualized in prevailing norms. Bias feedback must not undermine norms of egalitarianism but reinforce it. Further, feedback should not communicate that because bias is pervasive, it is normative or acceptable. Perceiving bias as normative could make egalitarian action less likely if it implies structural barriers that render individual action moot. Personal culpability, in this instance, would be trivial in comparison to the societally shared norm that bias is an acceptable part of human nature.

To counteract this potential impact of the feedback, it needs to suggest that while bias is pervasive, it is not acceptable, and hence not normative. And that while common, this is not a structural barrier because it is also commonly rejected. Shared norms of egalitarianism need to be communicated along with shared facts about the pervasiveness of bias. One could argue that the implication of a structural barrier could be avoided entirely by simply ignoring any mention of the pervasiveness of implicit bias. Merely communicate the egalitarian norm as part of the feedback. However, we feel that information about the pervasiveness of bias is essential to be communicated to circumvent the third condition that causes bias feedback to be demotivating.

2.3. Bias feedback that does not induce anxiety and arousal

Bias feedback can lead to the inference that the feedback recipient is being labelled as an abject moral failure with an unredeemable character. Such a label is highly undesirable and anxiety-provoking. To avoid such an inference, a delicate balance must be struck. On the one hand, the feedback about one's bias must be interpreted as a failure, one sufficient to motivate personal change in attitudes and behavior. However, the failure must not be so threatening to suggest one is, unavoidably, irredeemable and morally deficient. The best way to undercut the inference that feedback about bias implies a unique immorality is to communicate that such bias is common and pervasive, a natural part of human cognition. As discussed above, this attempt to soothe the threat to personal morality cannot simultaneously suggest that one need not be concerned about bias, or that bias is so common as to be normative, and thus trivial.

Thus, we propose that for bias feedback to be successful at initiating bias regulation it must balance these countervailing forces – communicating about the pervasiveness of bias to minimize a sense of personal *immorality*, while simultaneously engendering a sense of personal *responsibility* and culpability (triggering a shared egalitarian goal) that will not normalize bias. While balancing these two forces, it should simultaneously allow feedback recipients to develop a sense of efficacy over the control of bias. Interventions that meet these standards should be more likely to facilitate awareness of bias that persists over time, reduces defensiveness, and promotes prejudice-regulation. In a set of experiments we provide evidence for these hypotheses.

3. The present research

The goal of the current research is to evaluate the effectiveness of the proposed intervention in reducing defensive responding and, as a result, indirectly increasing bias awareness. Across four samples ($N = 1489$), and while utilizing multiples measures of bias awareness and forms of providing bias feedback, we find clear and convergent evidence that the proposed intervention reduces defensive responding and, as a result, directly and indirectly increases bias awareness. The proposed intervention contains features designed to (a) reduce perceived moral threat of implicit bias feedback, and (b) increase a sense of efficacy in the ability to minimize the impact of implicit bias on behavior, by emphasizing how it is malleable and subject to control. We investigate the effectiveness of this intervention across four experiments and a meta-

analysis of four independent samples, two of which were contacted for a follow-up. To test our hypotheses, we experimentally manipulated exposure to the intervention and receipt of bias feedback followed completion of a test of implicit racial attitudes.

3.1. Hypotheses

We tested four hypotheses in this research. Implicit bias feedback (vs. no feedback) is expected to heighten defensive responding, which manifests as viewing the feedback and the source of the feedback with reduced credibility and increased derogation (Hypothesis 1). This prediction is consistent with what has already been observed by prior work in this domain (e.g., Howell et al., 2013). More importantly, the intervention is expected to reduce defensive responding to implicit bias feedback, relative to a control group that received feedback but is not treated by the intervention (Hypothesis 2). These control conditions—no feedback and feedback without an intervention–provide a baseline for gauging the effectiveness of the intervention. Hypothesis 1 and 2 explore reduced defensive responding, but this intervention is also expected to indirectly increase bias awareness. Thus, defensive responding should mediate the effect of the intervention (vs. no intervention) on bias awareness at baseline (Hypothesis 3). Finally, it was also expected that the effect of the intervention on bias awareness would persist approximately 6-months post-feedback (Hypothesis 4).

3.2. Overview

In all experiments, our main objectives were to measure defensive reactions following exposure to the intervention or bias feedback. The bias feedback is always a *deception* of the experimental design – the feedback is not accurate and our goal is simply to have participants believe it is accurate. To accomplish this goal, we have them take an Implicit Association Test (IAT; Greenwald, Poehlman, Uhlmann, & Banaji, 2009) and we provide them with *bogus information* of this test having universal acceptance among the scientific community as a highly valid measure of implicit bias. Our concern is not whether the IAT actually has these properties, but to use it as a mere tool for distributing feedback about bias in a way that participants will take to heart. While the feedback is bogus and claims about the universal appeal of the test are exaggerated, participants are led to believe they are receiving accurate feedback from an accepted scientific tool. The validity of utilizing the IAT as a paradigm for manipulating beliefs about one's own attitudes and social cognition was first demonstrated by Vitriol, Reifen Tagar, Federico, and Sawicki (2019). Further, even though the feedback provided to participants is bogus, they did indeed complete the IAT, and we are able to compute and use their scores as control variables and interaction terms in various analyses below.

For each experiment, a large online sample of participants was recruited from Amazon's Mechanical Turk platform to complete a 20-min survey. Although Mturk samples are not a representative, random sample of the American public, Mturk samples are older and more diverse than typical samples of university students, and more nationally representative than typical internet samples (e.g., Berinsky, Huber, & Lenz, 2012). By utilizing Mturk, we are able to obtain a large, non-random sample of White Americans for each experiment with sufficient variability on demographic characteristics and, more importantly, the constructs of interest (see Paolacci & Chandler, 2014, on the usefulness of Mturk for psychological research). See Table 1 for all participant demographics.

Table 1
Demographic Characteristics of Participants from Study 1–4, Longitudinal Study.

Variables	Study 1	Study 2	Longitudinal	Study 3	Study 4
Sample Size	478	263	183	754	1004
Female, %	57.5%	63%	61%	64.4%	67.4%
Age Mean (SD)	37.46 (13.02)	35.59 (13.93)	40.89 (13.86)	35.15 (11.63)	37.27 (12.20)
Income >50 K, %	60.38%	52.47%	54.32%	50%	52.34%
< Bachelor's Degree, %	73.22%	68.83%	79.51%	83.95%	85.47%

Study 1 employed a single independent variable design (Feedback Only vs. Intervention) in which all participants received the same bias feedback, but half of the participants were randomly assigned to the Intervention (i.e., told prior to the feedback that bias is common and controllable) condition. Study 2 adopted the same design as Study 1, but also include a No Feedback condition (i.e., told nothing about their implicit attitudes) to establish a baseline for defensiveness. Study 1 and 2 utilized a feedback paradigm that characterized the results of a test of implicit racial attitudes in a way that both exaggerates the degree of implicit bias indicated by the results of the test and overstates the evidence for the predictive utility of implicit measures (but see Kurdi et al., 2019). As a result, this feedback heightens threat and therefore the motivation to reject it compared to the kind of feedback more typical to implicit tests and in prior studies examining defensive responding (e.g., Howell et al., 2013; Perry et al., 2015), as was used in Studies 3 and 4.

Thus, Study 1 and Study 2 provide a conservative test of our hypotheses—if the intervention is able to reduce defensive responding to harsh feedback, it is likely to be successful for less threatening forms of bias feedback. Participants who were enrolled in Study 1 or Study 2 and were assigned to either the Feedback Only or Intervention conditions were re-contacted 6-months after the original experiment for a follow-up survey that re-evaluated defensive reactions and bias awareness (i.e., Longitudinal Study). This allowed for an examination of the extent to which the cross-sectional effect of the Intervention persisted beyond the immediate measurement context.

Studies 3 and 4 were designed to "unpackage" the intervention examined in Studies 1 and 2, and to address methodological issues in our feedback paradigm and with our measurement of bias awareness. This was done by manipulating perceived efficacy and moral threat orthogonally to explore whether or not these components operate best in conjunction, or if it depends, instead, on the extent to which participants feel efficacious and able to control their bias (Dasgupta, 2013; Lai et al., 2014) or feel less morally threatened by the bias feedback. We also use a validated measure of bias awareness (Perry et al., 2015) and a feedback paradigm typical of implicit tests and utilized in prior studies examining defensive responding (e.g., Howell et al., 2013).

We end with a meta-analysis of the estimated effect of the Intervention compared to both control groups across Studies 1–4, which provides strong support for our hypotheses. Specifically, across four samples ($N = 1489$), using multiples measures of bias awareness and bias feedback paradigm, we find consistent evidence that, compared to the Feedback Only condition, the Intervention significantly reduced defensive responding and increased bias awareness both directly and indirectly (via reductions in defensive responding).

4. Study 1

4.1. Design

Experiment 1 *employed a single independent variable design* (Feedback Only vs. Intervention) in which all participants received the same bias feedback. An additional experimental condition was run concomitantly with the other conditions described above but was designed to test hypotheses different from what is addressed here. Information about this condition can be found in the supplemental materials. We report all measures used in this analysis here, and provide the exact language used for all items in supplemental materials, including a measure of affect not assessed here.

5. Method

5.1. Participants

Participants were 478 White U.S. citizens recruited from Amazon Mturk (57.5% females; mean age = 37.46, *SD* = 13.02). Most participants report a family income greater than 50 K (60.38%) and have earned at least a Bachelor's degree (73.22%). G*Power was used to determine the sample size needed to obtain adequate statistical power to detect mean level differences between the experimental and control group for medium effect sizes, and then Mturk participants were oversampled to adjust for the inclusion of non-Whites in the sample. Because estimated sample size was determined before any data analysis, it was not increased after preliminary data analyses. With the current sample size, it was estimated that the study had 64% power to detect a Cohen's *d* of 0.2 and 99% power to detect a Cohen's *d* of 0.5 or higher, for mean-level differences between conditions.

5.2. Procedure

Participants were recruited for a study of "Attitudes About People". The study advertised that it was primarily looking to recruit white U.S. citizens for a study, and that they would be compensated for their time.

Participants first viewed a consent form and were randomly assigned to experimental condition. Participants then proceeded to complete an IAT that they were told would measure "unconscious racial attitudes". The test presented them with pictures of men they needed to categorize according to race, and words they needed to categorize as good or bad, with accuracy and speed supposedly being measured for the purpose of yielding a "bias score" that would later be reported to them (Greenwald et al., 2009). This test was not actually used to provide the feedback, as it was merely a ruse to provide participants with a basis for feedback that was, in reality, randomly manipulated. That is, this test was used, in all experiments, as a false-feedback paradigm (e.g., Vitriol et al., 2019), in which participants were randomly assigned to receive bias feedback after treatment in the intervention condition (vs. control). The actual validity of the IAT as a measure is irrelevant to our purpose.

After completing the test, but *before* receiving any feedback, *only* participants assigned to the intervention condition received the following information, which serves as the intervention:

It is important to understand that this test does NOT *guarantee* that you are racially biased, nor does it mean that you have discriminated against racial minorities in the past. While unconscious racial bias is extremely common and is quite normal, it is also something that, once you are aware of it, you are able to control. For example, this test has been administered to a very large sample of people in countless different studies. The results from these studies indicate that the overwhelming majority of people harbor unconscious racial bias- even among people who strongly support racial equality and value racial tolerance. However, people who were made aware of their implicit bias were also better able to control it and minimize its influence on their judgment and behavior.

Social and behavioral scientists agree that unconscious preferences for some racial groups are a normal, basic feature of human cognition, and it has reliably been observed across most cultures and historical periods. In fact, one study determined that even social scientists who study racial discrimination commonly harbor unconscious racial bias. Most psychologists believe that unconscious beliefs, like the beliefs measured by this test, reflect the information available in the social environment and not some deep-rooted bigotry or hatred towards people in society. In this sense, unconscious racial bias is a basic feature of human cognition. It is a common and normal consequence of living in modern times, but it also something that people are able to control, once they become aware that it is influencing their thoughts and behavior.

Participants then completed a series of reading comprehension questions, which were intended to assess attention to, and accurate understanding of, the content of the intervention. After completing the intervention, participants in the intervention condition were provided with feedback. Participants assigned to the feedback condition did not undergo pre-feedback treatment, but received the same feedback as the intervention condition following completion of the "test of bias". The language and stimuli used for the exaggerated feedback is available in the supplemental materials.

Thus, comparisons between the two feedback conditions allow for a direct test of the effects of the intervention. Finally, participants completed the post-manipulation measures (described below), before being fully debriefed.

5.3. Measures

Table 2 provide the M(SD), alphas, and intercorrelations of all measures include in this analysis.

5.3.1. Manipulation checks

Participants answered 4 true or false items designed to measure comprehension of the information contained in the intervention. Such items include, "According to psychological scientists, unconscious racial bias is a basic feature of human cognition", "Most psychological scientists agree that people are able to control unconscious racial bias", "Prior research indicates that even social scientists who study race relations harbor unconscious racial bias", and "Unconscious beliefs reflect the information in the social environment, and not some deep-rooted bigotry or hatred towards racial minorities". Higher values represent a larger number of correct responses.

5.3.2. Defensive responding

Consistent with prior research examining resistance to self-threatening information (Kunda, 1987; Sherman, 2013), counter-attitudinal messages (Tormala & Petty, 2004; Vitriol et al., 2020) and bias feedback (Howell et al., 2013; Howell et al., 2017; Howell et al., 2017), we operationalize defensiveness here as derogation of the source of implicit bias feedback. Accordingly, we used original items designed for the purposes of this study, in which participants reported their belief in the validity, credibility, and objectivity of the test on which the feedback was based using a 7-point scale (1 = Not at all, 7 = Extremely"). These items include 1) "In your opinion, how credible is this test?", 2) "In your opinion, how objective is this test?", 3) "In your opinion, how valid are the results of this test? 4) "In your opinion, how useful is this test for understanding people's racial attitudes?" Responses were scaled such that higher values represent higher levels of defensiveness to the feedback.

5.3.3. Bias awareness

Participants reported the extent to which they perceive themselves as biased. 13-items measured participants' recognition of their own implicit racial bias and its social consequence. On a 7-point scale, participants responded to such items as, "How likely is it that your unconscious beliefs are unfavorable toward racial minorities?", "Do you believe that your unconscious racial attitudes influence your behavior towards racial minorities in an unfair way?", and "How likely is it that unconscious racial attitudes biases people's judgments and behavior towards racial minorities?" Higher values represent increased bias awareness.

5.3.4. Demographics

Participants reported their age, gender, race, family income, and level of education.

6. Results and discussion

Study 1 tests hypothesis 2 and 3 (that the interventions will reduce defensive responding and indirectly increase bias awareness), which

Table 2
Mean(SD), alphas, and correlations between all variables used in analyses for Study 1.

Variables	*M*	*SD*	α	1	2	3	4	5	6	7	8
1. Age	37.46	13.02	–	–							
2. Gender	0.43	0.50	–	0.04	–						
3. Income	3.13	3.03	–	0.04	0.06	–					
4. Education	5.44	2.41	–	−0.03	0.06	0.00	–				
5. Manipulation Check	0.91	0.21	0.60	0.08	−0.03	**0.18	0.01	–			
6. Defensive Responding	4.89	1.55	0.91	**0.13	0.02	0.05	0.04	**-0.14	–		
7. Bias Awareness	4.05	1.06	0.90	† − 0.08	*-0.10	−0.02	0.02	**0.28	**-0.55	–	
8. Race IAT D-Scores	0.39	0.40	–	**0.13	0.02	0.08	−0.02	0.08	−0.06	0.03	–

Higher values correspond with higher levels of the construct.
† $p < .10$.
* $p < .05$.
** $p < .01$.

involves a comparison between the intervention and the Feedback Only conditions. Implicit racial attitudes were included as covariates for *all* analyses in Study 1. While the inclusion of control variables that covary with our constructs strengthens our confidence in the independence and robustness of our theorized effects, model estimates without covariates included do not statistically nor substantively differ then estimates with covariates included. Only participants who completed the entire study were included in analyses. All other measures, manipulations, and exclusions are otherwise fully reported.

6.1. Manipulation checks

First, we evaluated whether participants attended to, comprehended, and retained the content of the intervention. To do so, we compared participants in the Intervention (M = 0.96, SD = 0.12) and Feedback Only (M = 0.83, SD = 0.28) conditions. Results of an independent-sample t-test indicates that participants in the former were significantly more likely to accurately comprehend and report judgments consistent with the content of the intervention ($t(371) = 5.79$, $p < .001$, 95% CI (0.08, 0.16). The supplemental materials also reports the result of a pilot study conducted on an independent sample ($N = 220$) that are consistent with these results.

6.2. Intervention will decrease defensive responding (h2) and indirectly increase bias awareness (H3)

We find strong and consistent support for both H2 and H3. The effect of experimental condition on defensive responding was significant, $F(1, 476) = 4.09$, $p = .044$, *Cohen's* d = 0.19. Participants in the intervention condition ($M = 4.74$, $SD = 1.54$) reported less defensive responding than participants in the Feedback Only condition ($M = 5.03$, $SD = 1.55$). Next, we examined whether the relationship between Intervention (vs. Feedback Only) and bias awareness was mediated by defensiveness, using the bootstrap-based method recommended by Preacher and Hayes (2004), in which 5000 bootstrap-replications were used to estimate confidence intervals. With defensiveness submitted as a mediator, the indirect effect of the Intervention (vs. Feedback Only) obtained significance on bias awareness ($b = 0.11$, $SE = 0.05$, (95% CI = 0.01, 0.21), $p = .04$). The direct effect of the Intervention (vs. Feedback Only) did not obtain significance on bias awareness ($b = 0.06$, $SE = 0.08$, (95% CI = −0.09, 0.22), $p = .43$). Furthermore, the total effect of the Intervention (vs. Feedback Only) did not reach conventional levels of statistical significance on bias awareness ($b = 0.17$, $SE = 0.10$, (95% CI = −0.02, 0.36), $p = .08$).

Thus, the results of Study 1 indicate that the intervention is effective at directly reducing defensive responding and, consequently, indirectly increases bias awareness. Importantly, these findings suggest that awareness of the commonality of implicit bias need not reduce culpability for bias and trivialize the issue. When appropriately framed as a transgression and a shortcoming relative to egalitarian goals that one can efficaciously address, people accept culpability. Though we do not explore it here, such culpability has previously been shown to trigger guilt and motivate prejudice regulation (e.g., Monteith et al., 2002). Our interpretation of these results do not substantively change when the measure of implicit racial attitudes[2] is included as an interaction term or is not included as a covariate in this analysis. Fig. 1 graphically represents mean-level defensive responding across condition.

7. Study 2

7.1. Design

The results of Study 1 indicate that the intervention was successful in both reducing defensive responding and increasing awareness of one's personal bias. Given concerns about the replicability of established effects in psychology (e.g., Open Science Collaboration, 2015), it is critical to determine whether a similar pattern of results can be observed on an independent sample. Study 2 provides a replication of the hypothesis that the intervention will reduce defensiveness and, consequently, increase bias awareness. It additionally includes a No Feedback control group to establish a baseline for defensiveness (a condition where a defensive response would not make any sense). Participants were randomly assigned to one of three conditions: No Feedback (i.e., told nothing about their implicit bias), Feedback Only (i.e., told they have implicit racial bias), or Intervention (i.e., told prior to the feedback that bias is common and controllable). In Study 2, as in Study 1, we used a bogus feedback paradigm in which participants were randomly assigned to receive feedback (with or without the pre-feedback intervention) or no feedback.

7.2. Participants

Participants were 263 White U.S. citizens recruited from Amazon Mturk (63% females; mean age = 35.59, $SD = 13.93$). Most participants report a family income greater than 50 K (52.47%) and have earned at least a Bachelor's degree (68.83%). G*Power was used to determine the sample size needed to obtain adequate statistical power to detect mean-level differences between the experimental and control group for medium effect sizes, and then Mturk participants were oversampled to adjust for the inclusion of non-Whites in the sample. Because estimated sample size was determined before any data analysis, it was not increased after preliminary data analyses. With the current sample size, it was estimated that the study had 38% power to detect a Cohen's *d* of 0.2 and 96% power to detect a Cohen's *d* of 0.5.

[2] Implicit attitudes were computed following the recommendations of Greenwald, Nosek, and Banaji (2003), with higher values representing increased automatic preference for White people relative to Black people.

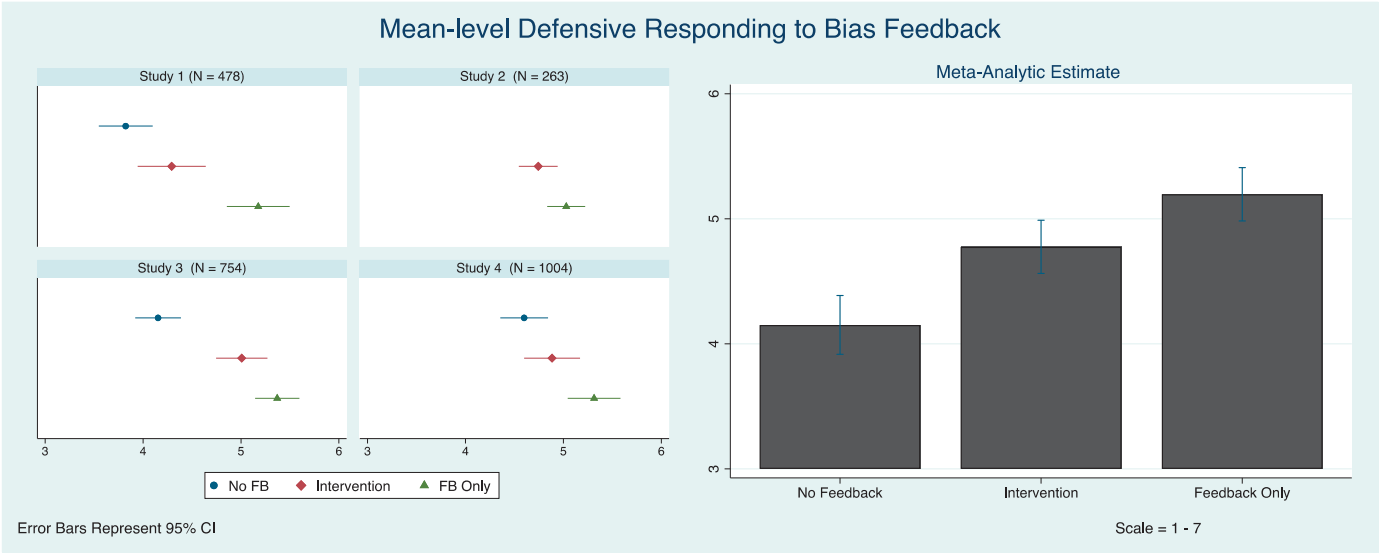

Fig. 1. Mean defensive responding across experiment. Error bars represent 95% CI.

7.3. Measures & procedure

Participants were recruited for a study of "Attitudes About People". The study advertised that it was primarily looking to recruit White U.S. citizens and would compensate participants $0.50 for their time. Study 2 employs the same procedure as Study 1, except no manipulation check measures were administered and the No Feedback condition was included. Furthermore, unlike Study 1, a battery of measures of individual differences were administered in Study 2 (measures used in this analysis are described below, whereas measures not used in this analysis are reported in the supplemental materials). The same dependent variables from Study 1 are administered in Study 2. All other measures, manipulations, and exclusions are otherwise fully reported. Table 3 provides the M(SD), alphas, and intercorrelations of all measures include in this analysis.

The following measures are unique to Study 2, administered before participants completed the test of "unconscious racial attitudes" and are used as covariates in analyses below. All of the results remain substantively and statistically the same without the covariates. Nonetheless, by including competing predictors of defensive responding in the same model estimating the effect of our experimental condition, our confidence in the robustness of our observations increases.

7.3.1. Social dominance orientation (SDO)

Participants complete the social dominance orientation scale (version 6) (Sidanius & Pratto, 2001), which serves as our measure of support for existing social hierarchy, and consists of 16 items to which the participants are asked to state their degree of agreement on a 7-point scale to such items as, "Some groups of people are simply inferior to other groups." Higher values represent higher levels of social dominance orientation.

7.3.2. Racial resentment (RR)

The racial resentment scale measures participants' explicit belief that blacks are unable or unwilling to work hard enough to overcome obstacles to success and are therefore undeserving of assistance or special favors (Kinder & Sanders, 1996). This is our measure of explicit racial attitudes. Participants responded to 4-items on a 5-point scale (1 = disagree strongly, to 5 = agree strongly), such as "It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites". Higher values represent higher levels of racial resentment.

7.3.3. Internal and external egalitarian motivations (IMS, EMS)

Participants reported the extent to which they are internally and

Table 3
Mean(SD), alphas, and correlations between all variables used in analyses for Study 2.

Variables	M	*SD*	α	1	2	3	4	5	6	7	8	9	10	11	12
1. Age	35.59	13.93	–	–											
2. Gender	0.36	0.48	–	†0.11	–										
3. Income	5.13	1.26	–	*0.15	0.04	–									
4. Education	5.76	3.18	–	*0.03	*-0.04	**0.22	–								
5. SDO	2.36	0.99	0.91	−0.01	**-0.18	−0.08	*0.13	–							
6. RR	3.05	1.08	0.73	**0.19	−0.01	−0.10	0.03	**0.49	–						
7. IMS	7.22	1.84	0.89	0.04	0.10	**0.20	−0.01	**-0.65	**-0.36	–					
8. EMS	4.16	2.05	0.84	† − 0.12	−0.02	0.07	0.00	*0.15	*0.15	−0.05	–				
9. SSS	3.64	0.88	0.86	*-0.14	−0.03	0.09	0.10	**-0.30	**-0.25	**0.17	−0.05	–			
10. Defensive Responding	4.42	1.58	0.93	0.05	−0.04	†0.12	0.01	−0.08	−0.02	†0.12	**-0.22	† − 0.11	–		
11. Bias Awareness	4.24	1.00	0.88	*-0.13	†0.11	0.06	−0.04	−0.10	**-0.27	−0.04	**0.17	*0.15	**-0.40	–	
12. Race IAT D-Scores	0.37	0.38	–	0.10	*0.13	−0.08	0.06	**0.17	**0.17	*-0.15	0.03	0.04	−0.08	0.06	–

SDO = social dominance orientation; RR = racial resentment; IMS / EMS = Internal or External Egalitarian Motivations; SSS = skepticism of social science. Higher values correspond with higher levels of the construct.

† $p < .10$.

* $p < .05$.

** $p < .01$.

externally motivated to control their prejudice (Plant & Devine, 1998). Internal motivations were measured across 5 items on 9-point scale (1 = strongly disagree to 9 = strongly agree), including such items as "I am personally motivated by my beliefs to be unprejudiced toward Black people". External motivations were also measured across 5 items on a 9-point scale (1 = strongly disagree to 9 = strongly agree), including such items as "I attempt to appear nonprejudiced toward Black people in order to avoid disapproval from others". Higher values on both scales correspond with increased internal or external egalitarian motivations.

7.3.4. Skepticism about social science (SSS)

Participants reported their general attitudes towards social science across 4-items, adapted from McCright, Dentzman, Charters, and Dietz (2013). Participants responded on a 5-point scale (1 = completely distrust, 5 = completely trust) to such items as, "How much do you distrust or trust social scientists to create knowledge that is unbiased and accurate?". Higher values represent more favorable opinions of social science.

8. Results and discussion

Study 2 tests Hypotheses 1–3. Only participants who completed the entire study were included in analyses. All other measures, manipulations, and exclusions are otherwise fully reported. Analyses were conducted in Study 2 the same way as for Study 1. For *all* analyses reported here, implicit and explicit racial attitudes, internal and external egalitarian motivations, SDO, and skepticism about social science are included as covariates. These measures were assessed to address research questions not examined in this paper. Without the covariates in the model, our observations and conclusion remain statistically and substantively unchanged.

8.1. Bias feedback will increase (H1), intervention will decrease (H2) defensive responding

This analysis investigates the hypothesis that Feedback Only (vs No Feedback) will *increase* defensive responding (H1) and the Intervention (vs. Feedback Only) will *reduce* defensive responding, using a one-way between-subjects ANOVA.

Consistent with H1 and H2, these results indicate a significant effect of experimental condition on defensive responding, $F(2, 254) = 21.97$, $p < .001$. Post hoc analyses using Duncan's method indicated that defensive responding was lower in the No Feedback ($M = 3.82$, $SD = 1.34$), compared to the Feedback Only condition ($M = 5.18$, $SD = 1.51$; 95% *CI for Mean Difference* (−1.86, −0.97), $p < .001$, *Cohen's d = 0.98*) and the Intervention condition ($M = 4.29$, $SD = 1.58$; 95% *CI for Mean Difference* (−1.02, −0.16), $p = .008$, *Cohen's d = 0.33*) conditions. Additionally, defensiveness was significantly reduced in the Intervention, compared to the Feedback Only, condition (95% *CI for Mean Difference* (−1.26, −0.38), $p < .001$, *Cohen's d = 0.59*).

Thus, consistent with prior research, bias feedback increased defensive responding (e.g., Howell et al., 2013; Perry et al., 2015). However, when the intervention preceded the bias feedback it reduced defensive responding. These effects emerged while controlling for implicit and explicit racial attitudes, social dominance orientations, and egalitarian motivations. Additional analyses were conducted to evaluate the extent to which implicit or explicit racial attitudes moderated the effect of experimental condition on defensive responding. Specifically, we examined the extent to which Racial Resentment or Race IAT-D scores moderated the effect of condition on defensive responding across a range of different model specifications. We conducted this analysis by testing the interaction between condition and each moderator, separately or simultaneously, with or without covariates included. Across all tests, the interaction did not obtain significance ($ps > 0.5$), indicating that the impact of bias feedback and the pre-feedback intervention did not vary across individual differences in racial attitudes. Fig. 1 graphically represents mean-level defensive responding across condition.

8.2. Defensive responding will mediate the effect of intervention on bias awareness (H3)

Here, we examine the hypothesis that defensive responding would mediate the effect of the Intervention (vs. Feedback Only) on bias awareness (H3) in the same way as for Study 1. The results of this analysis provide strong support for H3. With defensiveness submitted as a mediator, the indirect effect of the Intervention (vs. Feedback Only) obtained significance on bias awareness *(b = 0.27, SE = 0.09, (95% CI = 0.09, 0.44), p* = .002). The direct effect of the Intervention (vs. Feedback Only) did not obtain significance on bias awareness *(b = 0.12, SE = 0.14, (95% CI = −0.16, 0.39), p* = .40). Furthermore, the total effect of the Intervention (vs. Feedback Only) obtained significance on bias awareness *(b = 0.39, SE = 0.16, (95% CI = 0.06*, 0.71*), p* = .019). By decreasing defensive responding, the intervention indirectly increased bias awareness.

9. Longitudinal study

The results of Study 1 and 2 demonstrate that the intervention can reduce defensive responding to implicit bias feedback, and, as a result, increase awareness of bias in the self and others. However, these effects were observed cross-sectionally, that is, immediately after experimental treatment—the extent to which these effects persist (or not) beyond the immediate context remains unknown. By reducing defensive responding, this intervention may help facilitate strategies that aid in both the recognition and regulation of prejudice responding that extend beyond the experimental context. For example, research has identified "cues for control" as fundamental to the self-regulation of prejudiced-responding (see Monteith, Mark & Ashburn-Nado, 2010). This self-regulatory process entails the development of associations between the consequences of prejudiced-responding (i.e., guilt) and contextual stimuli that triggered the response, which improves individuals' ability to inhibit prejudicial responses over time (Monteith & Mark, 2009). Cues for control motivate prejudice-regulation because individuals have become aware of the propensity for prejudiced-responding and its consequences. One implication of this dynamic is that strategies that induce short-term increases in bias awareness—by, for example, reducing defensive responding– may help individuals to recognize bias in themselves and others over time.

9.1. Participants and procedure

In the Longitudinal Study, between December 28th 2016 and January 22nd 2017, participants in the Feedback Only and Intervention conditions of Study 1 (May 25th to June 25th 2016) & Study 2 (June 25th to July 30th 2016) were re-contacted approximately six months after the original experimental session and were re-administered measures of defensive responding and bias awareness. Additional measures administered at that time, but not relevant to the current study, assessed participants' political party and ideological identification, voting behaviors for the 2016 presidential election, and attitudes and evaluations of contemporary political issues and actors. Of the 648 eligible participants (Study 1 $n = 478$, Study 2 $n = 170$), all were contacted but only 183 were retained (27%; Study 1 $n = 111$, Study 2 $n = 72$). Participants who did or did not return for the follow-up survey did not significantly differ in mean-levels of defensiveness, bias awareness, or implicit racial attitudes ($ps > 0.05$), and the demographic characteristics are highly similar (see Table 1). Further, the rate of attrition among participants in the Intervention condition (26%) was comparable to that for participants in the Feedback Only condition (27%). The Longitudinal Study sample included 183 White U.S. citizens recruited from Amazon Mturk (61% female; mean age = 40.89, $SD = 13.86$). Most participants report a

family income greater than 50 K (54.32%) and have earned at least a Bachelor's degree (79.51%). With this sample size, it was estimated that the Longitudinal Study had 37% power to detect a Cohen's *d* of 0.2 and 95% power to detect a Cohen's *d* of 0.5 or higher.

10. Results

This design allows for an analysis of the extent to which the indirect cross-sectional effects of the feedback intervention conditions on bias awareness persist 6-months later (H4). This test involves a comparison between the intervention group and the Feedback Only condition. Because these samples were recruited consecutively, and because the follow-up survey was administered concurrently for all eligible participants, a dummy-variable to represent the sample from Study 1 or 2 was not included, although doing so does not change the interpretation of the results. Furthermore, we also include as a covariate implicit racial attitudes, although omitting that from our model does not change our results. These analyses are available upon request. Participants from Studies 3 and 4 were not re-contacted and are therefore not included in this analysis.

First, the effect of the intervention on defensive responding and bias awareness, measured approximately six months post-feedback, was assessed using an independent-sample *t*-test. These analyses indicate that, compared to the Feedback Only condition ($M = 3.56$, $SD = 1.13$), the Intervention condition ($M = 4.00$, $SD = 1.03$) had significantly higher levels of bias awareness *(*$t(181) = 2.76$, $p = .006$, *Cohen's* d = 0.42) six months after feedback. However, differences in defensiveness were not observed in the follow-up survey ($t(181) = 1.23$, $p = .22$, *Cohen's* d = 0.19; Feedback Only $M = 4.14$, $SD = 1.42$; Intervention $M = 3.88$, $SD = 1.39$). Thus, the intervention led to mean-level increases in bias awareness approximately six months removed from the receipt of feedback, suggesting that this effect of the intervention persisted beyond the original experimental context.

Next, we examined whether the cross-sectional reduction in defensive responding not only increased bias awareness in the short-term (H3), but also accounts for the persistence of higher levels of bias awareness at the follow-up. To test this, we again conducted mediation analysis using the bootstrap-based method recommended by Preacher and Hayes (2004), in which 5000 bootstrap-replications were used to estimate confidence intervals. With baseline defensiveness submitted as a mediator, the indirect effect of the Intervention (vs. Feedback Only) condition was marginally significant for bias awareness at the 6-month follow-up (*b = 0.17, SE = 0.09, (95% CI* = −0.00, 0.33*)*, $p = .055$). The direct effect of the Intervention (vs. Feedback Only) was significantly associated with increased bias awareness *(b = 0.29, SE = 0.14, (95% CI = 0.01, 0.56))*, $p = .041$). However, the total effect of the Intervention (vs. Feedback Only) did not obtain significance on bias awareness *(b = 0.45, SE = 0.16, (95% CI = 0.14, 0.77)*, $p = .005$).

Thus, by reducing defensive responding in the short term, the intervention led to increases in bias awareness that were stable approximately six months later. These findings are consistent with other research investigating the longitudinal effect of bias interventions (Devine et al., 2012), and suggest that this intervention may have helped facilitate strategies that aid in both the recognition and regulation of prejudice-responding that extends beyond the experimental context. Future research should examine this hypothesis more directly using existing methods for studying prejudice-regulation (see Monteith & Mark, 2009).

11. Study 3

11.1. Design

Study 3 manipulated perceived efficacy and moral threat as separate dimensions of the intervention, administered manipulated checks for the psychological constructs targeted by the intervention, utilized a validated measure of bias awareness (Perry et al., 2015), and provided feedback in a less exaggerated or condemnatory way. Accordingly, Study 3 employed a 1 (No Feedback Control) + 2 (Efficacy; High vs. No Information) x 2 (Moral Threat: Low vs. No Information) design. The condition with no information is equivalent to the Feedback Only condition, and the condition with High Efficacy + Low Moral Threat is equivalent to the Intervention condition. We expect that, consistent with the results of Studies 1–2, participants exposed to an Intervention (vs. Feedback Only) that increases both kinds of efficacy and perceptions of the commonality of bias in the general population will reduce defensive responding and, consequently, increase bias awareness (Hypotheses 2–3). We also explore the effects of each dimension of the intervention, compared to the control groups, independently, but do not advance any a priori hypotheses. In Study 3, as in Study 1 and 2, we again used a bogus feedback paradigm.

12. Method

12.1. Participants

Participants were 754 White U.S. citizens recruited from Amazon Mturk (64.4% females; mean age = 35.15, $SD = 11.63$). Most participants report a family income greater than 50 K (50%) and have earned at least a Bachelor's degree (83.95%). G*Power was used to determine the sample size needed to obtain adequate statistical power to detect mean level differences between each experimental and control group for medium effect sizes, and then Mturk participants were oversampled to adjust for the inclusion of non-Whites in the sample. Because estimated sample size was determined before any data analysis, it was not increased after preliminary data analyses. With the current sample size, it was estimated that the study had at least 52% power to detect a Cohen's *d* of 0.2 and 99% power to detect a Cohen's *d* of 0.5 or higher.

12.2. Procedure

Participants were recruited for a study of "Attitudes About People". The study advertised that it was primarily looking to recruit White U.S. citizens. As before, the name of the study was intended to increase the expectation that one's beliefs and attitudes about other people would be directly measured.

Participants first viewed a consent form for the study and were then randomly assigned to experimental conditions. All participants completed the "test of bias". However, unlike in studies 1 and 2, participants in study 3 were provided the following instructions prior to the IAT:

In this study, you will be shown pictures of several individuals, and will be asked to pair each of these pictures with a list of words. Please remember that we are interested in your perspective. There are no right or wrong answers to the questions, and your first response is usually the best. After completing the task you will be asked to provide some information on your opinion about current events and your basic demographics information. This is a challenging task, but it's necessary for the aim of this study. Please try hard to help us in our analysis.

After completing the "test of bias", and prior to feedback, some participants were administered the intervention. The text for the Low Moral Threat prompt included the following information:

Unconscious racial bias is extremely common in the general population. Scientists agree that unconscious preferences for some racial groups are a feature of human cognition, and this bias has been reliably observed across most cultures and historical periods. In fact, one study determined that even social scientists who study racial discrimination commonly harbor unconscious racial prejudice. Most psychologists believe that unconscious racial bias is a basic feature of human cognition.

The text for the High Efficacy prompt included the following information:

Fortunately, it is possible for people to become more aware of their unconscious biases. It is also possible for people to become aware of how it is influencing the way they are thinking and acting. With practice, awareness is possible to achieve. Furthermore, when people are aware of their unconscious biases, they can control it and change how they treat racial minorities. It might even be possible to transform one's unconscious racial bias so that it ultimately disappears.

Participants in the High Efficacy, Low Moral Threat condition were presented with both prompts simultaneously; participants in the Feedback Only condition were not provided with any information at this stage. After exposure to the intervention, participants were then provided with the following information.

The results of your test will be on the next screen once it has been computed. Please click "Next" to review the results once it appears. This could take about a minute.

The results of this test indicate that you have a moderate automatic preference in favor of White relative to Black people.

Participants in the experimental condition than proceeded to complete the dependent measures. In contrast, participants in the no-feedback control group proceeded straight to the dependent measures. Finally, participants answered questions about their demographics and were fully debriefed.

We utilized the same measure of defensive responding in Study 3 as for Studies 1–2. Experiment 3 included novel measures of manipulation checks and an established measure of bias awareness, described below. The exact language of all the measures is available at the end of the appendix. Only participants who completed the entire study were included in analyses. All other measures, manipulations, and exclusions are otherwise fully reported.

12.3. Measures

We used the same measure of defensiveness in Study 3 as for Study 1 and 2. Below we describe the new manipulation check and bias awareness measures. We report all measures used in this analysis here, and provide the exact language used for all items in supplemental materials, including a measure of affect not assessed here. Table 4 provide the M(SD), alphas, and intercorrelations of all measures include in this analysis.

12.3.1. Manipulation checks

Participants responded to a series items designed to evaluate perceived efficacy and moral threat. Perceived efficacy was measured with two items on a 7-point scale (1 = strongly agree to 7 strongly disagree): (1) "Because racial bias is unconscious, most people can never see it in themselves" and (2) "Even people who are aware of their unconscious racial bias often fail to minimize its influence on their judgment and behavior.". Perceived moral threat was measured with two items on a 7-point scale (1 = strongly agree to 7 strongly disagree): (1) "Unconscious racial bias is common in the American population" and (2) "Most psychological scientists agree that unconscious racial bias is a basic feature of human cognition." Higher values on both scales correspond with increased perceptions of efficacy and commonality of bias or less moral threat.

12.3.2. Bias awareness

Participants responded to Perry et al. (2015)'s bias awareness measure. For this scale, participants responded to 4-items on a 7-point scale ranging from strongly disagree to strongly agree. Example items included: "When talking to Black people, I sometimes worry that I am unintentionally acting in a prejudiced way", and "Even though I know it's not appropriate, I sometimes feel that I hold unconscious negative attitudes toward Blacks". Higher values were coded to indicate higher levels of bias awareness.

13. Results and discussion

In Study 3, we completed two sets of analysis. First, we seek to replicate the results of Study 1 and 2 by comparing the Intervention condition (i.e., High Efficacy, Low Moral Threat) to the No Feedback and Feedback Only conditions. This confirmatory test was conducted using a one-way between-subjects ANOVA to evaluate the effects of the intervention on defensive responding, compared to the control groups. Second, we also explore the independent effect of each feature of the intervention– the high efficacy only condition, and the low moral threat only condition – to the Feedback Only condition. For this exploratory analysis, a one-way between-subjects ANOVA was also conducted to compare the impact of efficacy and moral threat, independently and in conjunction, against the control groups. Post-hoc pairwise comparisons for both sets of analyses were conducted using Duncan's method. For *all* analyses, when implicit racial attitudes are included as a covariate, the results do not change. For this reason, we omit control variables from analyses in Study 3 and 4 to simplify the presentation of our results, having included these covariates in our analysis of Study 1 and 2, although doing so does not change our estimates or conclusion.

13.1. Manipulation check

Before proceeding with our confirmatory and exploratory analysis concerning variability in defensive responding across conditions, we first examine the effect of our efficacy and moral threat manipulation on the manipulation check items. For this analysis, separate analyses were performed for each factor, such that the effects of the high efficacy (vs. no information) on the efficacy manipulation check and low moral threat (vs. no information) on the moral threat items, were conducted independently using a one-way between-subjects ANOVA. In support of the validity of the efficacy and threat manipulation, we obtain marginally significant effects of the efficacy factor on perceived efficacy ($F(1, 601) = 3.19$, $p = .07$; High Efficacy $M = 4.62$, $SD = 1.43$; No Efficacy Info $M = 4.41$, $SD = 1.45$) and marginally significant effects of the moral threat factor on perceived moral threat ($F(1, 597) = 2.83$, $p = .09$; Low Moral Threat $M = 5.31$, $SD = 1.41$; No Moral Threat Info $M = 5.11$,

Table 4
Mean (SD), alphas, and correlations between all variables used in analyses for Study 3.

Variables	M	*SD*	α	1	2	3	4	5	6	7	8	9
1. Age	35.15	11.63	–	–								
2. Gender	1.64	0.48	–	*0.08	–							
3. Income	6.02	3.06	–	**0.11	0.02	–						
4. Education	5.10	1.36	–	**0.12	0.03	**0.29	–					
5. MC Efficacy	4.64	1.45	0.68	*-0.08	0.02	0.00	0.03	–				
6. MC Moral Threat	5.30	1.41	0.73	**-0.12	0.01	0.00	†0.07	**0.53	–			
7. Defensive Responding	4.92	1.55	0.91	*-0.08	*-0.07	−0.02	−0.02	**0.37	**0.39	–		
8. Bias Awareness	3.81	1.48	0.82	*-0.08	0.00	0.03	**0.11	**0.42	**0.46	**-0.44	–	
9. Race IAT D-Scores	0.41	0.39	–	0.06	0.00	0.05	−0.05	0.03	−0.01	0.02	−0.02	–

Note. †$p < .10$. *$p < .05$. **$p < .01$. MC efficacy / moral threat = manipulation check to evaluate perceived efficacy to control or commonality of bias (i.e., moral threat), respectively. Higher values correspond with higher levels of the construct.

SD = 1.46). Furthermore, high levels of efficacy and low levels of moral threat were significantly associated with reduced levels of defensiveness (Efficacy, b = −0.23, CI 95% (−0.32, −0.15), $p < .001$; Moral Threat, b = −0.30, CI 95% (−0.38, −0.21), $p < .001$) and increased levels of bias awareness (Efficacy, b = 0.24, CI 95% (0.17, 0.32), $p < .001$; Moral Threat, b = 0.35, CI 95% (0.28, 0.43), $p < .001$). We did not observe a significant effect of the moral threat factor on perceived efficacy or of the efficacy factor on perceived moral threat (*ps* > 0.29). These observations emerged in regression analysis when including both sets of manipulation check items as predictors in the same model, and with implicit racial attitudes as covariates. Together, these results provide support for the validity of our manipulation and our expectation that these constructs are important and consequential for defensive responding and bias awareness.

13.2. Bias feedback will increase (H1) defensiveness; intervention decreases (H2) defensiveness, indirectly increase bias awareness (H3)

Here we seek to confirm our observations from Study 1 and Study 2 by comparing the intervention condition (i.e., High Efficacy, Low Moral Threat) to the two control groups, No Feedback and Feedback Only conditions. Consistent with the results of Study 1 and 2, we again find strong support for H1, H2 and H3. The effect of experimental condition on defensive responding was significant, $F(2, 442) = 26.15$, $p < .001$, *Cohen's* d = 0.70. Post hoc analyses using Duncan's method indicated that defensive responding was lower in the No Feedback condition ($M = 4.15$, $SD = 1.45$), compared to the Feedback Only ($M = 5.37$, $SD = 1.37$; *95% CI for Mean Difference* (−1.19, −0.52), $p < .001$, *Cohen's d = 0.86*) and the Intervention ($M = 5.01$, $SD = 1.62$; *95% CI for Mean Difference* (−1.58, −0.86), $p < .001$, *Cohen's d = 0.56*) conditions. Additionally, defensiveness was significantly reduced in the Intervention, compared to the Feedback Only condition (*95% CI for Mean Difference* (−0.70, −0.02), $p = .037$, *Cohen's d = 0.24*). This comparison between the Intervention and Feedback Only produced a small effect size (the implications of which are discussed more in the meta-analysis and discussion section).

Next, we examined the extent to which defensiveness mediated the effect of the Intervention (vs. Feedback Only) on bias awareness. With defensiveness submitted as a mediator, the indirect effect of Intervention (vs. Feedback Only) obtained significance on bias awareness (*b = 0.13, SE = 0.07, (95% CI = 0.00, 0.27), p* = .046). The direct effect of the Intervention (vs. Feedback Only) did not obtain significance on bias awareness (*b = 0.04, SE = 0.15, (95% CI = −0.26, 0.34), p* = .41), nor was the total effect of the Intervention (vs. Feedback Only) on bias awareness (*b = 0.17, SE = 0.17, (95% CI = −0.15, 0.49), p* = .304).

In sum, the results of Study 3 replicate the results of Study 1 and 2 and provide additional support for H1, H2 and H3. Fig. 1 graphically represents mean-level defensive responding across condition.

13.2.1. Exploratory analysis of independent effects of high efficacy and low moral threat

Next, we explore the independent effect of each feature of the intervention by comparing the remaining set of conditions to the Feedback Only control conditions. This analysis was conducted using a one-way between-subjects ANOVA that tests for differences across all 5 conditions. Because this is an exploratory analysis, we did not apply post-hoc adjustments to correct for family-wise error. Accordingly, we interpret as evidence against the null any effect between the Feedback Only control and experimental condition observed at $p < .10$. The effect of experimental condition on defensive responding was significant, $F(4, 748) = 13.67$, $p < .001$, *Cohen's* d = 0.41. Compared to the Feedback Only condition ($M = 5.37$, $SD = 1.37$), defensive responding was marginally lower in the high efficacy ($M = 5.06$, $SD = 1.55$; *95% CI for Mean Difference* (−0.66, 0.03), $p = .075$, *Cohen's d = 0.21*) and significantly lower in the low moral threat conditions ($M = 5.03$, $SD = 1.51$; *95% CI for Mean Difference* (−0.68, −0.01), $p = .049$, *Cohen's d = 0.24*), and the Intervention condition (*95% CI for Mean Difference* (−0.71, −0.02), $p = .039$, *Cohen's d = 0.23*). These results were marginally significant but suggest that high efficacy and low moral threat can operate independently to reduce defensive responding but appear to work most effectively when administered in conjunction. In Experiment 4, we seek to clarify this dynamic by again manipulating moral threat and efficacy independently.

Participants showed a significant reduction in defensiveness in the Intervention condition, *95% CI for Mean Difference* (−0.83, −0.03), $p = .036$, compared to the Feedback Only condition. No other comparison approached $p < .10$.

14. Study 4

14.1. Design

In Study 4, we seek to replicate the results of Studies 1, 2, and 3, and extend our investigations by manipulating levels of perceived moral threat (high vs. low vs. no information) and levels (high vs. low vs. no information) of two distinct conceptions of efficacy (self vs. response-efficacy). Efficacy beliefs may concern the perceived effectiveness of a behavioral response for attaining a specified goal (i.e., response-efficacy; e.g., "I can control the influence of implicit bias on my judgment once I am aware of it") or the perceived ability to engage in the behavioral response (i.e., self-efficacy; "I can become aware of the influence of implicit processes on my judgment"; Bandura, 1977, 1982; Rogers, 1975; Witte & Allen, 2000). Because our existing intervention affirmed both self and response-efficacy, in addition to reducing perceived moral threat, this design allows us to further disentangle the features of this intervention and determine which, together or apart, drive the reduction in defensive responding and increases in bias awareness that was observed in the previous studies. Table 5 provides descriptive information about the features included in each experimental condition. In Study 4, as in Study 1, 2, and 3, we used a bogus feedback paradigm.

15. Method

15.1. Participants

Participants were 1005 White U.S. citizens recruited from Amazon Mturk (67.4% females; mean age = 37.27, *SD* = 12.20). Most participants report a family income greater than 50 K (52.34%) and have earned at least a Bachelor's degree (85.47%). Given the number of conditions included in this design, we targeted approximately 100 participants per cell, and then Mturk participants were oversampled to adjust for the inclusion of non-Whites in the sample. Because estimated sample size was determined before any data analysis, it was not increased after preliminary data analyses. With the current sample size, it was estimated using G*Power that the study had at least 41% power to detect a Cohen's *d* of 0.2 and 97% power to detect a Cohen's *d* of 0.5 or higher.

15.2. Measures and procedure

Study 4 employed the same procedure and same measure of manipulation checks, defensive responding, and bias awareness as in Study 3 (see Table 6 for information about each condition in Study 4). Furthermore, Study 4 adopted the same language used in Study 3 for (a) providing bias feedback, and (b) to manipulate perceived increased efficacy and reduced moral threat. The exact language used in pre-feedback prompt for each experimental condition is provided in the supplemental materials. All other measures, manipulations, and exclusions are otherwise fully reported (with the exception of a measure of affect, which is described in the supplemental materials).

Table 5
Mean (SD), alphas, and correlations between all variables used in analyses for Study 4.

Variables	M	*SD*	α	1	2	3	4	5	6	7	8	9
1. Age	37.26	12.20	–	–								
2. Gender	1.67	0.47	–	*0.07	–							
3. Income	6.02	3.05	–	**0.11	−0.03	–						
4. Education	5.12	1.36	–	**0.13	0.03	**0.35	–					
5. MC Efficacy	4.32	1.59	0.65	*-0.07	0.00	−0.03	† − 0.06	–				
6. MC Moral Threat	4.92	1.33	0.75	† − 0.06	0.02	0.00	**0.09	**0.42	–			
7. Defensive Responding	5.10	1.44	0.91	†0.06	†0.06	*0.06	*0.08	**-0.27	**-0.38	–		
8. Bias Awareness	3.62	1.48	0.81	−0.05	0.04	0.04	**0.16	**0.20	**0.40	**-0.37	–	
9. Race IAT D-Scores	0.36	0.39	–	**0.14	0.01	−0.01	−0.02	*0.07	0.04	−0.05	0.02	–

Note. †$p < .10$. *$p < .05$. **$p < .01$. MC efficacy / moral threat = manipulation check to evaluate perceived efficacy to control or commonality of bias (i.e., moral threat), respectively. Higher values correspond with higher levels of the construct.

Table 6
Condition Assignment, Study 4.

Condition	Feedback		Moral Threat			Self-Efficacy			Response-Efficacy		
	Yes	No	High	Low	No Info	Yes	No	No Info	Yes	No	*No Info*
1		**Y**			**Y**			**Y**			**Y**
2	**Y**				**Y**			**Y**			**Y**
3	**Y**		**Y**					**Y**			**Y**
4	**Y**		**Y**			**Y**			**Y**		
5	**Y**		**Y**			**Y**				**Y**	
6	**Y**		**Y**				**Y**			**Y**	
7	**Y**			**Y**				**Y**			**Y**
8	**Y**			**Y**		**Y**			**Y**		
9	**Y**			**Y**		**Y**				**Y**	
10	**Y**			**Y**			**Y**			**Y**	

16. Results and discussion

In Experiment 4, we test Hypothesis 2 and 3, by comparing the Intervention (i.e., High Response-efficacy, High Self-efficacy, and Low Moral Threat) condition to the Feedback Only and No Feedback conditions. This confirmatory test was conducted using a one-way between-subjects ANOVA. Similarly, we explore the independent effect of each feature of the intervention by comparing those conditions to the Feedback Only condition using a one-way between-subjects ANOVA. Post-hoc pairwise comparisons for both sets of analyses were conducted using Duncan's method. For Hypothesis 3, we examine the role of defensive responding in mediating the relationship between the Intervention (vs. Feedback Only) on bias awareness.

16.1. Manipulation check

First, we again examine the effect of our efficacy and moral threat manipulation on the manipulation check items. For this analysis, separate analyses were performed for high vs. low levels of each feature, such that the effects of the condition with both kinds of efficacy (vs. no response-efficacy and no self-efficacy) on the efficacy manipulation check and low moral threat (vs. high moral threat) on the moral threat items, were conducted independently using a one-way between-subjects ANOVA. In support of the validity of the efficacy and threat manipulation, we obtain marginally significant effects for the efficacy factor on perceived efficacy ($F(1, 396) = 5.26$, $p = .022$; Yes Efficacy $M = 4.49$, $SD = 1.49$; No Efficacy $M = 4.12$, $SD = 1.57$) and significant effects for the moral threat factor on perceived moral threat ($t(1, 805) = 33.16$, $p < .001$; Low Moral Threat $M = 4.64$, $SD = 1.45$; High Moral Threat Info $M = 5.18$, $SD = 1.21$). We did not observe a significant effect of the moral threat factor on perceived efficacy or of the efficacy factor on perceived moral threat ($ps > 15$).

To strengthen our confidence in the claim that our intervention successfully targeted perceived efficacy and commonality of bias, we compared the effect of the Intervention on the manipulation check against both control groups. We conducted the same analysis as above.

In support of the validity of the intervention, we obtain significant effects for perceived efficacy ($F(2, 296) = 3.42$, $p = .034$; Intervention, $M = 4.47$, $SD = 1.46$; Feedback Only, $M = 4.01$, $SD = 1.50$; No Feedback, $M = 4.53$, $SD = 1.51$) and for moral threat ($F(2,300) = 6.64$, $p = .002$; Intervention, $M = 5.37$, $SD = 1.09$; Feedback Only, $M = 4.77$, $SD = 1.31$; No Feedback, $M = 5.10$, $SD = 1.09$). Post hoc analyses using Duncan's method indicated that perceived efficacy was significantly higher in the Intervention, compared to Feedback Only (*95% CI for Mean Difference* (0.04, 0.88), $p = .032$, *Cohen's d = 0.*56) but not statistically different from the No Feedback (*95% CI for Mean Difference* (−0.46, 0.35), $p = .78$, *Cohen's d = 0.*20) conditions. Similarly, post hoc analyses using Duncan's method indicated that perceived moral threat was significantly lower in the Intervention, compared to Feedback Only (*95% CI for Mean Difference* (0.26, 0.95), $p < .001$, *Cohen's d = 0.*56) and marginally significantly lower then the No Feedback (*95% CI for Mean Difference* (−0.04, 0.58), $p = .09$, *Cohen's d = 0.*20) conditions.

Furthermore, when both perceived efficacy and moral threat were included in the same model as predictors for the entire sample of participants, high levels of efficacy were associated with significantly reduced levels of defensiveness (b = −0.12, CI 95% (−0.18, −0.06), p < .001) but was not significantly related to bias awareness (b = 0.04, CI 95% (−0.01, 0.10), $p = .14$). Reduced perceptions of moral threat predicted a significant reduction in defensiveness (b = −0.35, CI 95% (−0.42, −0.28), p < .001) and a significant increase in bias awareness (b = 0.342, CI 95% (0.35, 0.49), p < .001). These observations emerged in regression analysis when including both sets of manipulation check items as predictors in the same model, and with implicit racial attitudes as covariates.

Together, this analysis provides additional support for the validity of our manipulation and our expectation that these constructs are important and consequential for defensive responding and bias awareness.

16.2. Bias feedback will increase (H1) defensiveness; intervention decreases (H2) defensiveness, indirectly increase bias awareness (H3)

Here we seek to confirm our observations from Studies 1–3 by

comparing the Intervention condition (i.e., High Self-Efficacy, High Response-Efficacy, Low Moral Threat) to the two control groups, No Feedback and Feedback Only conditions. Consistent with the results of Studies 1–3, we again find strong support for H1, H2, and H3. The effect of experimental condition on defensive responding was significant, $F(2, 300) = 6.85$, $p = .001$, *Cohen's* d = 0.42. Post hoc analyses using Duncan's method indicated that defensive responding was lower in the No Feedback ($M = 4.60$, $SD = 1.27$), compared to the Feedback Only ($M = 5.31$, $SD = 1.29$; *95% CI for Mean Difference* (−1.12, −0.31), $p < .001$, *Cohen's d = 0.56*), but not in the Intervention ($M = 4.88$, $SD = 1.43$; *95% CI for Mean Difference* (−0.65, 0.08), $p = .13$, *Cohen's d = 0.20*) conditions. Thus, the intervention reduced defensiveness to baseline levels. Most importantly, defensiveness was significantly reduced in the Intervention condition, compared to the Feedback Only, condition (*95% CI for Mean Difference* (−0.82, −0.05), $p = .027$, *Cohen's d = 0.31*). Fig. 1 graphically represents mean-level defensive responding across condition. This comparison between the Intervention and Feedback Only condition produced a small-to-medium sized effect (the implications of which are discussed more in the meta-analysis and discussion section).

Next, we examined the extent to which defensiveness mediated the effect of the Intervention (vs. Feedback Only) on bias awareness. With defensiveness submitted as a mediator, the indirect effect of the Intervention (vs. Feedback Only) obtained significance on bias awareness *(b = 0.20, SE = 0.10, (95% CI = 0.01, 0.40), p = .038)*. The direct effect of the Intervention (vs. Feedback Only) did not obtain significance on bias awareness *(b = 0.08, SE = 0.01,* (95% CI = −0.29, 0.48), $p = .67$). Furthermore, the total effect of the Intervention (vs. Feedback Only) did not obtain significance on bias awareness *(b = 0.28, SE = 0.21, (95% CI = −0.61, −0.33), p = .18)*.

16.2.1. Exploratory analysis of independent effects of high efficacy and low moral threat

Next, we explore the independent effect of each feature of the intervention – self-efficacy, response efficacy, and perceptions of moral threat – by comparing all conditions to the Feedback Only condition. This analysis was conducted using a one-way between-subjects ANOVA that tests for differences across all conditions, and with post-hoc pairwise comparisons with the Feedback Only condition. Because this is an exploratory analysis with 8 pairwise comparisons, we did not apply post-hoc adjustments to correct for family-wise error. Accordingly, we interpret as evidence against the null any effect between the Feedback Only control and experimental condition observed at $p < .10$. The effect of experimental condition on defensive responding was significant, $F(9, 995) = 5.13$, $p = .007$, *Cohen's* d = 0.30. Participants showed a significant reduction in defensiveness in the Intervention condition, *95% CI for Mean Difference* (−0.83, −0.03), $p = .036$, compared to the Feedback Only condition. No other comparison approached $p < .10$.

Thus, we find evidence in support of the effectiveness of an intervention that increases both self-efficacy and response efficacy, and reduces perceived moral threat—but, unlike in Study 3, we do not observe evidence to suggest that these features alone reduce defensiveness, nor do we find that reducing moral threat is effective when paired with low efficacy (or vice versa).

17. Meta-analysis of studies 1–4

Next, we conducted a meta-analysis across all of our studies ($N = 1489$), in which we examine differences in defensive responding across the intervention ($n = 571$) and control (No Feedback $n = 351$; Feedback Only $n = 567$) conditions using two approaches. For both sets of analyses, we report the results of models that do not include any covariates. Table 7 reports the M(SD) for the key dependent variables used in this study, separated by and aggregated across study.

Table 7
Mean (SD) for critical variables used in analyses for Study 1–4 + Meta-analysis.

	Mean (SD) x Study				
	Study 1 (*N* = 263)	Study 2 (*N* = 478)	Study 3 (*N* = 445)	Study 4 (*N* = 303)	Meta-Analysis (*N* = 1489)
Defensive Responding	4.22 (1.58)	4.89 (1.55)	4.84 (1.57)	4.91 (1.38)	4.80 (1.54)
Bias Awareness	4.24 (1.00)	4.05 (1.06)	3.81 (1.44)	3.56 (1.43)	3.93 (1.27)
Race IAT	0.37 (0.38)	0.39 (0.40)	0.40 (0.40)	0.35 (0.37)	0.38 (0.39)

Higher values correspond with higher levels of defensive responding, bias awareness, and pro-White implicit bias.

17.1. Bias feedback will increase (H1), intervention will decrease (H2) defensive responding

First, we conducted a meta-analysis with a fixed effects model (Goh, Hall, & Rosenthal, 2016). Results reveal that, across Studies 2, 3, & 4, defensive responding was lower in the No Feedback, compared to the Feedback Only (dCombined (95% CI) = 0.79 (0.64, 0.95), Zcombined = 9.87, pCombined <0.001) and Intervention condition (dCombined (95% CI) = 0.38 (0.23, 0.54), Zcombined = 4.99, pCombined<0.001). More importantly, however, across all 4 studies, we find that defensive responding is significantly lower in the Intervention, compared to the Feedback Only, condition (dCombined (95% CI) = 0.27 (0.14, 0.40), Zcombined = 4.15, pCombined <0.001).

Second, we conducted multilevel modeling with experiment as a random-intercept term and tested the effect of experimental condition on defensive responding, while treating between-experiment variability as a random effect. Table 8 summarizes the results of this analysis and Fig. 1 graphically represents the meta-analytic estimate of mean-level defensive responding across conditions. The results of the random-intercept model are consistent with the results of the fixed effects model; the intervention significantly reduced defensive responding to bias feedback. This is a substantively small-to-medium sized effect, comparable to the average effect size reporting in psychological research (Gignac & Szodorai, 2016), and has important and potentially large practical implications. As a benchmark, the observed effect of the intervention in reducing defensive responding to bias feedback is similar in impact as the effect size for the relationship between the use of a pain reliever and the alleviation of pain from a headache (Funder & Ozer, 2019).

17.2. Defensive responding will mediate the effect of intervention on bias awareness (H3)

Finally, we examine the hypothesis that defensive responding would mediate the effect of the Intervention (vs. Feedback Only) on bias awareness (H3). We first present the results of each experiment independently, and then report the meta-analytic effect. In order to account for clustering of responses within experiments in the 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, we performed a bootstrap analyses per recommendation by Preacher and Hayes (2004) with 5000 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).

For the meta-analytic estimate of the effects of the Intervention (vs. Feedback Only) on bias awareness, through defensive responding, the indirect effect obtained significance on bias awareness *(b = 0.16, SE = 0.04, (95% CI = 0.09, 0.23), p < .001)*. Furthermore, the direct effect of

Table 8
Multilevel regression models on defensive responding, with experiment as a random-intercept term.

Parameter	No FB (1) vs. FB Only (0)		No FB (1) vs. Intervention (0)		Intervention (1) vs. FB Only (0)	
	b (95% CI)	*SE*	*b* (95% CI)	*SE*	*b* (95% CI)	SE
Fixed						
Intercept	**5.22 (5.01, 5.43)	0.11	**4.76 (4.50, 5.01)	0.11	**5.19 (5.00, 5.38)	0.10
Condition	**-1.07 (−1.29)	0.11	**-0.58 (−0.80, −0.36)	0.13	**-0.42 (−0.59, −0.24)	0.09
Random						
var(Intercept)	0.03 (0.004, 0.10)	0.03	0.05 (0.01, 0.28)	0.04	0.02 (0.002, 0.16)	0.02
LR-Test	**5.58		**10.30		**4.02	

Note. †$p < .10$ *$p < .05$ **$p < .01$.

the Intervention (vs. Feedback Only) also obtained significance on bias awareness ($b = 0.06$, $SE = 0.01$, $(95\%\ CI = 0.05, 0.08)$, $p < .001$). Finally, the total effect was also significant ($b = 0.22$, $SE = 0.04$, $(95\% CI = 0.14, 0.31)$, $p < .001$). Thus, the results of this analysis support our hypothesis that, by decreasing defensive responding, the intervention indirectly and directly increased awareness of and culpability for personal bias.

18. Discussion

Do interventions and workshops regulate individual-level bias? Are people receptive to learning about bias? Does education about the many subtle forms of bias having unintended effects on how one thinks and acts (for a review, see Banaji & Greenwald, 2016; Bargh, 2017) lead to the desired effect? The path from negative feedback to bias reduction is fraught. Critically, there are times it has the opposite of the desired effect, triggering resentment, anger, denial, polarization, motivated reasoning, and backlash. Education about structural bias places the fault in poorly regulated social systems, whereas individual bias is personal. Well-intentioned programs that educate about implicit bias may antagonize and engender defensiveness and induce backlash and threat, rather than awareness and opportunities for growth.

These questions have gone largely unexplored in psychological science and industry, even as organizations strive to educate workers and improve conditions. Many proposed remedies lack empirical evidence (Paluck & Green, 2009). For those bias-reduction strategies that have been demonstrated to succeed, it is not clear that these can induce long-term change in attitudes and behavior (Lai et al., 2016; Stone, et al., 2020) or have the desired effects when introduced in field settings by non-experts (e.g., Redford & Ratliff, 2016). Ironically, in such cases where behavior change is most needed, pointing out that need can lead to motivated reasoning. Kunda (1987) showed that feedback about a behavior that posed a serious health threat did not lead to the engine of motivation being revved up to counteract the threat, but instead to one justifying the self-threatening behavior through derogation of the feedback and doubt regarding its veracity. Ditto and Lopez (1992) provided further evidence of motivated reasoning by illustrating the lengths to which perceivers will go to rationalize feedback that suggests their health may be threatened. Such behavior places people at serious health risk.

Individuals, organizations, and institutions committed to anti-discrimination principles would be remiss to ignore the empirical evidence on implicit racial bias and its relation to discriminatory outcomes (see Jost et al., 2009). Interventions using such principles share the assumption that increasing awareness of propensities towards racial bias (i.e., negative feedback) is an effective strategy for reducing its impact. But the possibility exists that such interventions worsen the situation by creating the defensiveness and backlash described above. In four experiments, we show how educational initiatives and anti-bias interventions can attenuate this kind of motivated reasoning and instead increase the kind of bias awareness needed for sustainable regulation and reduction in bias. By reducing perceived moral threat upon learning that one has implicit bias and increasing perceived efficacy in controlling bias, this intervention reliably reduced defensive responding and led to stable mean-level differences in bias awareness approximately 6-months after the experimental session, relative to baseline.

The longitudinal effects of the Intervention on bias awareness suggests that providing feedback in a way that attenuates defensive responding may also stimulate the development of self-regulatory cues for controlling prejudiced-responding (e.g., Monteith, Mark, & Ashburn-Nardo, 2009). In contrast, implicit racial bias feedback that activates defensive responding may lead to negative associations with the source of the feedback, instead of with prejudiced-responses, thereby failing to instigate self-regulatory processes. Unfortunately, the development of cues for control as an account for the persistence of bias awareness is only indirectly evidenced from the current research design. Future research should examine this hypothesis more directly using established paradigms and methods for studying prejudice-regulation (see Monteith & Mark, 2009).

Despite the strength of our evidence across studies, our research nonetheless has several limitations. First, in our exploratory analysis in Studies 3 and 4, we lack adequate statistical power to reliably observe a difference between the control groups and conditions that included independent features of the Intervention. Those studies were primarily designed to replicate the effects of the Intervention that was observed in Studies 1 and 2, which was successful, but nonetheless may not have been diagnostic of whether specific features of the intervention were primarily driving these effects. As it stands, we conclude that the Intervention, as a package, is reliably successful in achieving its goals. But while we observe that features of that Intervention can be impactful when administered independently, additional investigations are needed in order to confidently make this determination.

Additional research should continue to investigate other strategies that can help reduce defensive responding to implicit bias feedback and whether or not doing so can predict behavioral change and prejudice-reduction, in addition to promoting bias awareness. Our meta-analytic results indeed suggests that the Intervention is a small-to-medium sized effect, which yields important practical implications (Gignac & Szodorai, 2016). However, the Intervention does not consistently reduce defensive responding to baseline levels, that is, to the levels observed in the No Feedback condition (see results of the meta-analysis). The impact of bias feedback is larger than the Intervention, suggesting that there may be additional measures needed to eliminate defensive responding completely. Similarly, that our Intervention reduces but does not eliminate defensive responding also raises questions about whether or not it will translate to meaningful behavior change. The current studies examined bias awareness, but not actual behavior. Because bias awareness is a necessary, albeit insufficient, step for prejudice-regulation (Carter, et al., in press), we are confident our Intervention can help promote sustainable reductions in bias behavior (Moskowitz & Vitriol, 2021). Nonetheless, future research is needed to directly investigate the impact of this Intervention on other outcomes of interest, including stereotypic responding (e.g., Moskowitz & Li, 2011), and, perhaps most importantly, actual behavior in applied settings.

The generalizability of our observations is also limited by its exclusive reliance upon MTurk samples. MTurk samples may be older and

more diverse than student samples, and more nationally representative than typical internet samples (e.g., Berinsky et al., 2012), but are not a representative, random sample of the American public. Further, our sample was limited to U.S. citizens. Future research should investigate the generalizability of our observations to samples more representative of the U.S. and the international community (e.g., Vitriol et al., 2019). Additionally, whether increasing efficacy and reducing moral threat can reduce defensive responding to different modes of bias feedback (e.g., interpersonal confrontation; Czopp et al., 2006) or in regard to other targets of bias (e.g., women, LGBTQ), also remains unaddressed. Investigating the generalizability and boundary conditions of these findings will likely bear fruit for diversity science and bias education.

In sum, bias awareness is blocked when an individual dismisses valid feedback through rationalization and defensive responding. Our experiments 1) examine how interventions can be framed to mitigate the defensiveness and backlash associated with feedback about bias, and 2) illustrate that reductions in defensiveness can promote both short-term and long-term increases in awareness of bias, the theoretical cornerstone of prejudice regulation and egalitarian behavior.

Data availability

All of the data and data syntax are available at: https://osf.io/pbrdn/?view_only=b2070c097d7143f3917cb72215aa5c37

Author contributions

J.A. Vitriol is responsible for developing the initial project concept. Both authors were responsible for the theoretical framework, drafting the final manuscript, and the study design and data collection, analysis, and interpretation. Study 1, Study 2, and the Pilot Study reported in this paper was submitted by J.A. Vitriol to the Psychology Department at the University of Minnesota in partial fulfillments of the requirements for a doctoral degree.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

The authors would like to thank Eugene Borgida, Mark Snyder, Christopher Federico, and Howard Lavine for their help in developing the study concept; Mahzarin R. Banaji and members of her lab, Dominic Packer, Adam Magerman, Naomi Rothman, Alexandera Sackett, and Jose Causadias for critical comments and feedback on early drafts and iterations of this project; Andrew Sell and Alicia Hofelich Mohr at the University of Minnesota for help with software design; Center for the Study of Political Psychology at University of Minnesota, Lehigh University Faculty Research Grant, and the Society for Personality and Social Psychology for supporting this research program.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2021.104165.

References

Ajzcn, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179–211.
Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology, 94*, 60–74.
Amodio, D. M., & Hamilton, H. K. (2012). Intergroup anxiety effects on implicit racial evaluation and stereotyping. *Emotion, 12*, 1273–1280.
Axt, J. R., Casola, G. M., & Nosek, B. A. (2018). Reducing social judgment biases may require identifying the potential source of bias. *Personality and Social Psychology Bulletin, 45*(8), 1232–1251.
Banaji, M. R., & Greenwald, A. G. (2016). *Blindspot: Hidden biases of good people*. Delacorte Press.
Bandura, A. (1977). Self-efficacy: Towards a unifying theory and the organization. *Psychological Review, 84*(2), 191–215.
Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37* (2), 122–147. https://doi.org/10.1037/0003-066X.37.2.122.
Bandura, A. (1989). Self-regulation of motivation and action through internal standards and goal systems. In L. A. Pervin (Ed.), *Goal concepts in personality and social psychology*. Hillsdale, N.J: Erlbaum.
Bandura, A. (1991). Self-regulation of motivation through anticipatory and self-reactive mechanisms. In R. A. Dienstbier (Ed.), *Nebraska symposium on motivation (Vol. 38). Lincoln, NE: Univ*. Nebraska Press.
Bargh, J. (2017). *Before you know it: The unconscious reasons we do what we do*. New York, NY: Touchstone.
Baumeister, R. F., & Campbell, W. K. (1999). The intrinsic appeal of evil: Sadism, sensational thrills, and threatened egotism. *Personality and Social Psychology Review, 3*, 210–221.
Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical Turk. *Political Analysis, 20*(3), 351–368.
Carnes, M., Devine, P. G., Baier Manwell, L., Byars-Winston, A., Fine, E., Ford, C. E., … 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.
Carter, E. C., Onyeador, I. N., & Lewis, N. A., Jr. (2021). What do we know about (implicit) bias and what does it mean for bias reduction training? *Behavioral Science & Policy, 6*(1), 57–70.
Carver, C. S., & Scheier, M. F. (1998). *On the self-regulation of behavior*. New York: Cambridge University Press.
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.
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.
Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology, 47*, 233–279.
Daumeyer, N. M., Onyeador, I. N., Brown, X., & Richeson, J. A. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology*, 84.
Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48*(6), 1267–1278.
Devine, P. G., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology, 60*(6), 817.
Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63*, 568–584.
Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science, 11*, 319–323.
Forscher, P., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. (2019). A meta-analysis of change in implicit bias. *Psychological Bulletin, 117*(3), 522–559. https://doi.org/10.1037/pspa0000160.
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.
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*, 1611–1624.
Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Send and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–158.
Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78.
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.
Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216.
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.
Hansen, F. (2003). Diversity's business case: doesn't add up. *Workforce, 824*, 28–32.
Higgins, E. T., Strauman, T., & Klein, R. (1986). Standards and the process of selfevaluation: Multiple affects from multiple stages. In R. M. Sorrentino, & E. T. Higgins (Eds.), *Handbook of Motivation and Cognition: Foundations of Social Behavior* (pp. 23–63). New York: Guilford Press.
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.

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*, 714–720.
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.
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*, 2–21.
Howell, J. L., Redford, L., Pogge, G., & Ratliff, K. A. (2017). Responding defensively to IAT feedback. *Social Cognition, 35*(5), 520–562.
Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior, 29*, 39–69.
Kawakami, K., Dovidio, J., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotype associations on stereotype activation. *Journal of Personality and Social Psychology, 78*, 871–888.
Kenrick, A., Sinclair, S., Richeson, J. A., Versoksy, S., & Lun, J. (2016). Moving while black: Intergroup attitudes influence judgments of speed. *Journal of Experimental Psychology. General, 145*, 147–154.
Kinder, D. R., & Sanders, L. M. (1996). *Divided by color. Racial politics and democratic ideals*. Chicago: University Chicago Press.
Klinger, E. (1975). Consequences of commitment to and disengagement from incentives. *Psychological Review, 82*(1), 1–25.
Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research, 36*(2), 249–277.
Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology, 53*(4), 636–647.
Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., … Banaji, M. R. (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *The American Psychologist, 74*(5), 569–586.
Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., & Frazier, R. S. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*(4), 1765.
Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., … Nosek, B. A. (2016). Reducing implicit racial preferences: II, Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*, 1001–1016.
Lipman, J. (2018). How diversity training infuriates men and fails women. *Time.* Retrieved from https://time.com/5118035/diversity-training-infuriates-men-fails-women/.
McCright, A. M., Dentzman, K., Charters, M., & Dietz, T. (2013). The influence of political ideology on trust in science. *Environmental Research Letters, 8*(4), Article 044029.
Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin, 36*, 512–523.
Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice reduction efforts. *Journal of Personality and Social Psychology, 65*, 469–485.
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.
Monteith, M. J., & Mark, A. Y. (2009). The self-regulation of prejudice. *Handbook of Prejudice, Stereotyping, and Discrimination*, 507–523.
Monteith, M. J., Mark, A. Y., & Ashburn-Nardo, L. (2010). The self-regulation of prejudice: Toward understanding its lived character. *Group Processes & Intergroup Relations, 13*(2), 183–200. https://doi.org/10.1177/1368430209353633.
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*. UK: Taylor & Francis Routledge.
Moskowitz, G. B. (2002). Preconscious effects of temporary goals on attention. *Journal of Experimental Social Psychology, 38*(4), 397–404.
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.
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.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). aac4716.
Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology, 60*, 339–367.
Paolacci, G., & Chandler, J. (2014). Inside the turk understanding mechanical turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184–188.
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.
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.
Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Experimental Social Psychology, 75*(3), 811.
Plant, E. A., & Devine, P. G. (2003). The antecedents and implications of interracial anxiety. *Personality and Social Psychology Bulletin, 29*(6), 790–801.
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.
Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*(3), 369–381.
Redford, L., & Ratliff, K. A. (2016). Hierarchy-legitimizing ideologies reduce behavioral obligations and blame for implicit attitudes and resulting discrimination. *Social Justice Research, 29*(2), 159–185.
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.
Richeson, J. A., & Shelton, J. N. (2007). Negotiating interracial interactions: Costs, consequences, and possibilities. *Current Directions in Psychological Science, 16*, 316–320.
Richeson, J. A., & Trawalter, S. (2008). The threat of appearing prejudiced and race-based attentional biases. *Psychological Science, 19*, 98–102.
Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude Change1. *The Journal of Psychology, 91*(1), 93–114.
Schlenker, B. R., & Leary, M. R. (1982). Social anxiety and self-presentation: A conceptualization model. *Psychological Bulletin, 92*, 641–669.
Sekaquaptewa, D., Takahashi, K., Malley, J., Herzog, K., & Bliss, S. (2019). An evidence-based faculty recruitment workshop influences departmental hiring practice perceptions among university faculty. *Equality, Diversity and Inclusion: An International Journal, 38*(2), 188–210.
Shepperd, J., Malone, W., & Sweeny, K. (2008). Exploring causes of the self-serving bias. *Social and Personality Psychology Compass, 2*, 895–908.
Sherman, D. K. (2013). Self-affirmation: Understanding the effects. *Social and Personality Psychology Compass, 7*(11), 834–845. https://doi.org/10.1111/spc3.12072.
Sidanius, J., & Pratto, F. (2001). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.
Spencer, S. J., Fein, S., Wolfe, C. T., Fong, C., & Duinn, M. A. (1998). Automatic activation of stereotypes: The role of self-image threat. *Personality and Social Psychology Bulletin, 24*(11), 1139–1152.
Steele, C. M. (1988). The Psychology of Self-Affirmation: Sustaining the Integrity of the Self. *Advances in Experimental Social Psychology, 21*, 261–302.
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). NJ: Lawrence Erlbaum Associates, Inc.
Stone, J., Moskowitz, G. B., Zestcott, C., & Wolsiefer, K. (2019). *Testing active learning workshops for reducing implicit stereotyping of Hispanics by majority and minority group medical students*. Stigma and Health.
Tangney, J. P. (1995). Shame and guilt in interpersonal relationships. In J. P. Tangney, & K. W. Fischer (Eds.), *Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride (p. 114–139)*. Guilford Press.
Tormala, Z. L., & Petty, RE. (2004). Resistance to Persuasion and Attitude Certainty: The Moderating Role of Elaboration. *Personality and Social Psychology Bulletin, 30*(11), 1446–1457. https://doi.org/10.1177/0146167204264251.
Vitriol, J. A., Calanchini, J., & O'Shea, B. (2020). *Less Bias, Yet More Defensive: The Role of Control Processes*. Manuscript submitted for publication.
Vitriol, J. A., Reifen Tagar, M., Federico, C. M., & Sawicki, V. (2019). Ideological uncertainty and investment of the self in politics. *Journal of Experimental Social Psychology, 82*, 85–97.
Vorauer, J. D., & Kumhyr, S. M. (2001). Is this about you or me? Self-versus other-directed judgments and feelings in response to intergroup interaction. *Personality and Social Psychology Bulletin, 27*, 706–719.
Wicklund, R. A., & Gollwitzer, P. M. (1982). *Symbolic self-completion*. Hillsdale, NJ: Erlbaum.
Witte, K., & Allen, M. (2000). A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health Education & Behavior, 27*, 591–615. https://doi.org/10.1177/109019810002700506.
Wrosch, C., Scheier, M. F., Carver, S. C., & Schulz, R. (2003). The importance of goal disengagement in adaptive self-regulation: When giving up is beneficial. *Self and Identity, 2*, 1–20. https://doi.org/10.1080/15298860309021.