PREPRINT AHEAD OF PRINT

Defensive Responding to IAT Feedback

Jennifer L. Howell
Ohio University

Liz Redford
Kate A. Ratliff
Gabrielle Pogge
University of Florida

We investigated whether learning that one is biased predicts defensive reactions to feedback on Implicit Association Tests (IATs; Greenwald, McGhee, & Schwarz, 1998). In an archival data set (Study 1, N = 219,426) and in an online experiment (Study 2, N = 1,225), people responded most defensively to feedback when: (a) their implicit and explicit attitudes were more discrepant than congruent, (b) their implicit attitudes aligned more with societal bias than did their explicit attitudes (e.g., a preference for Straight People relative to Gay People), and (c) they were majority group members (e.g., White participants in a race-relevant task) rather than minority group members. Next, in an online experiment (Study 3, N = 418), we demonstrated that receiving feedback indicating one is biased causes greater defensiveness. In turn, greater defensiveness, predicts lower intentions to engage in egalitarian behavior.

A robust body of research supports the notion that ego-defense mechanisms protect the self from psychological threat (Baumeister, Dale, & Sommer, 1998). That is, when people encounter information that threatens their self-view, they may engage in defensive strategies like denying, downplaying, dismissing, or avoiding the information (Howell & Shepperd, 2012; McQueen, Vernon, & Swank, 2013; Pyszczynski & Greenberg, 1987). In doing so, people can create and sustain positive self-views even in light of evidence to the contrary (Dunning, 2007; Taylor & Brown, 1988).

For most people in the United States, a cherished view is seeing oneself as egalitarian (O'Brien et al., 2010). Americans often report having very little prejudice because: (a) they want to adhere to group norms (Crandall, Eshleman, & O'Brien, 2002), (b) they want to be fair and just toward others (Banaji & Bhaskar, 2000), and (c) they are unaware of their own prejudice (Dovidio & Gaertner, 2004). Thus, any suggestion that one is prejudiced, even implicitly, could be quite threatening and may induce self-protective strategies (Howell et al., 2013; Rudman, Dohn, & Fairchild, 2007). In fact, Arkes and Tetlock (2004) write that, “To call someone prejudiced or racist in early twenty-first century America is to comment on both the cognitive competence and moral standards of that individual” (p. 268). In the present paper we build on the idea that evidence of one's prejudice should provoke defensive responding. Specifically, we examine whether people display defensiveness in response to feedback about their personal implicit bias, especially when that feedback contradicts their self-reported beliefs.

The Implicit Association Test and Threat

One of the most widely-used implicit measures is the Implicit Association Test (IAT) (Greenwald, McGhee, & Schwartz, 1998). The IAT is a computerized task that assesses the

relative strength of associations between two target categories (e.g., insects and flowers) and two concepts (e.g., good and bad) by pairing each group with each concept (e.g., flowers and good; insects and bad; flowers and bad; insects and good) and using response latency to operationalize association strength. Although the IAT has been subject to both empirical and theoretical debate (e.g., Blanton, Jaccard, Gonzales, & Christie, 2006; Greenwald, Nosek, & Sriram, 2006), researchers use it frequently to measure a variety of implicit cognitions (Greenwald, Poehlman, Uhlmann, & Banaji, 2009).

One of the richest sources of data using the IAT is the Project Implicit website (https://implicit.harvard.edu). Visitors to the site can complete IATs on topics of their choosing and then receive feedback about their implicit attitudes. Since 2008, more than 10 million IATs have been completed on the Project Implicit site, and, in 2012, the site averaged around 11,000 completed tests each day. Each participant also reports their explicit attitudes, and those who receive feedback also answer questions about their opinions of the feedback. As such, Project Implicit provides a natural context to determine whether the discrepancy between explicit attitudes and feedback about implicit attitudes predicts defensive responding.

There are a variety of reasons why people may hold implicit attitudes that differ from their explicit views, and such discrepancy is quite common (Briñol, Petty, & Wheeler, 2006; Gawronski & Bodenhausen, 2011). For instance, most people favor their in-group more implicitly than explicitly (Axt, Ebersole, & Nosek, 2014; Howell, Gaither, & Ratliff, 2015). This implicit-explicit discrepancy can have a variety of consequences including producing cognitive dissonance (Rydell, McConnell, & Mackie, 2008), changing how people process attitude-relevant information (Briñol et al., 2006; Rydell et al., 2008; Shoda, McConnell, & Rydell, 2014), and even prompting self-defense (Jordan, Spencer, Zanna, Hoshino-Browne, & Correll, 2003). In the present study, we build on this earlier work showing implicit-explicit discrepancy influences cognitions to examine whether people who are directly made aware of such discrepancy may respond defensively to such information. Specifically, we examine participants' reactions to the feedback they received about their implicit bias from the Project Implicit website. In doing so, we do not intend to imply that implicit attitudes are somehow more "true" than explicit attitudes. Indeed, they clearly each have their own behaviorally-predicative power (Greenwald et al., 2009). Instead we aim only to understand the consequences of feedback suggesting divergence between implicit and explicit attitudes.

Our predictions about these reactions are drawn from self-verification theory, according to which people feel threatened when receiving feedback that is inconsistent with their self-views (Swann, 1990). This idea is rooted in theories of cognitive consistency (Abelson, 1968), which suggest that people experience cognitive dissonance when they learn information that violates their beliefs (Festinger, 1962). To the extent that explicit attitudes reflect one's endorsed self-perception (O'Brien et al., 2010), people may be threatened by implicit-explicit misalignment.

Of course, it is possible that people may not be threatened by implicit-explicit misalignment. Indeed, given that some are not aware of their implicit attitudes, they may not perceive themselves to be responsible for the content of those attitudes (Redford & Ratliff, 2015). Moreover, IAT feedback, particularly on the Project Implicit site, is presented as automatic and out of personal control, not as a "true" attitude. As such, it is possible that people do not respond defensively to IAT feedback. Still, prior research suggests that, at least in the context of the Black-White IAT, some people respond defensively.

In one experiment, White participants were told that a Black-White Race IAT measures either implicit racial bias or knowledge of a cultural stereotype. In the former condition, they display outcomes indicative of stereotype threat: they are subsequently more likely than in the latter conditions to display pro-White implicit bias (Frantz, Cuddy, Burnett, Ray, & Hart, 2004). Further, they are subsequently more likely to experience higher levels of implicit, but not explicit, self-esteem, indicating automatic self-esteem compensation (Rudman et al., 2007). Other research suggests that people naturally begin the Black-White IAT expecting that it will confirm their egalitarian self-view. As a result, when they read forewarnings that their feedback is likely to indicate they are implicitly prejudiced, they respond defensively by avoiding their feedback altogether (Howell et al., 2013).

Other research more directly supports the idea that people might respond defensively to IAT feedback. Most related to the present work is research by Monteith, Voils, and Ashburn-Nardo (2001) where participants completed the Black-White Race IAT then reported on their experience. Although participants did not receive feedback about their score, they were generally able to detect that the IAT would indicate pro-White bias, and they experienced negative affect as a result. Evidencing defensiveness, 63% of participants attributed their performance on the IAT to some factor irrelevant to racial bias (e.g., errors in measurement, color patterns). Similarly, an examination of six years of archival data from the Project Implicit website suggested that people derogated the Black-White Race IAT to the extent that their implicit feedback and explicit attitudes differed, and particularly when IAT feedback indicated they were more pro-white than they indicated explicitly (Howell, Gaither, & Ratliff, 2014). Thus, existing research suggests that people should respond defensively to feedback about their performance on an IAT. However, research thus far has focused mostly on the Black/White Race IAT, has rarely investigated responses to actual feedback, has primarily involved college student samples, and has not investigated defensiveness specifically. As such, earlier work does not thoroughly and experimentally investigate the question we pose here.

Overview of the Present Studies

Consistent with the notion that people want to receive feedback that is consistent with their self-views (Swann, 1990), we hypothesized that defensiveness would emerge to the extent that implicit and explicit attitudes differed. Further, consistent with the notion that people want to see themselves positively (Sedikides & Gregg, 2008)—in the present case, as egalitarian—we predicted that people would be more defensive to the extent that their implicit attitudes aligned with preferences for majority groups (e.g., White, straight) relative to minority groups (Black, gay). We also hypothesized that defensive responses to IAT feedback would predict reduced intentions to address personal anti-Black sentiment.

We examined these hypotheses in three studies. The first study used data collected on the Project Implicit research website from participants who completed one of nine IATs in 2012 (Stereotyping IATs: Black-White/Weapons-Harmless Objects, Gender/Career-Family, Gender/Science-Liberal Arts, Native-White/Foreign-American, Asian-White/Foreign-American; Evaluative IATs: Gay-Straight/Good-Bad, Young-Old/Good-Bad, Abled-Disabled/Good-Bad, Arab Muslim-Other People/Good-Bad). We investigated whether the discrepancy between explicit attitudes and implicit attitudes (the latter about which participants received feedback) predicted participant defensiveness, which we operationalized as discounting or downplaying the relevance of IAT feedback.

In the first study, we also examined whether majority group members (e.g., White, male, straight) would be more defensive than

minority group members (e.g., Black, female, non-straight). We were interested in whether interactions between these variables would emerge; specifically, whether majority group members who learned their implicit bias aligned more with societal stereotypes than their explicit attitude (e.g., that they preferred straight people to gay people more implicitly than explicitly) would be particularly defensive to the extent that their implicit and explicit attitudes differed.

In the second study, we expand on this initial archival analysis in several ways. First, we investigate whether the proposed effects hold across a variety of other defensiveness outcomes, including affective reactions and other forms of derogation. In doing so, we also provide a construct validation of the measure of defensiveness used in the initial study. We also investigate whether implicit-explicit discrepancy has indirect effects on other outcomes (e.g., behavioral intentions) via defensive responding.

Finally, in the third study, we examine whether defensiveness is not just predicted, but caused, by learning that one implicitly prefers a majority group over a minority group. We do so by manipulating Black-White Race IAT feedback. This allows us to examine whether White people who learn that they implicitly prefer White people to Black people, compared to White people who learn they prefer White and Black people equally, respond more defensively to IAT feedback, and whether such defensiveness leads to reduced intentions to address personal anti-Black sentiment. That is, Study 3 experimentally investigates the correlations suggested in Studies 1 and 2. In so doing, it represents the first experimental investigation into how people respond to learning that they implicitly favor White over Black people.

Theoretical and Practical Contribution

Although earlier studies suggest that people may respond defensively to IAT feedback by experiencing stereotype threat (Franz et al., 2004), engaging in automatic self-esteem enhancement (Rudman et al., 2007), avoiding the feedback (Howell et al., 2013), derogating the feedback (Howell et al., 2014), or blaming outside forces (Monteith et al., 2001), the present work extends this prior work in five important ways. First, prior studies investigating IAT-related defensiveness focused exclusively on responses to the Black-White Race IAT; in the current research we examine responses to feedback regarding a variety of attitude objects, making a stronger case for the generality of the defensiveness processes we investigate. Second, this is only the second line of inquiry to directly test whether people respond defensively to feedback about their actual performance on a measure of implicit attitudes, and the first to do so across several attitude objects. Indeed, most focus on how people react to the process of taking the IAT or to contrived feedback. Third, this is the first experimental investigation into how people's defensive responses to IAT feedback are related to their intentions to engage in egalitarian behavior. Fourth, the present studies expand on earlier work by investigating a variety of defensiveness outcomes across a diverse set of stimuli. Fifth, we investigate theoretically important nuance in predictors of responses to IAT feedback, including aspects of implicit-explicit discrepancy (i.e., magnitude, direction) and demographic characteristics of the participants completing the task (i.e., group membership).

Study 1

Study 1 examined whether implicit-explicit discrepancy and group membership predicted defensive responding to IAT feedback in an ecologically valid context using archival data analysis.

Study 1: Method

Participants

Participants were 219,426 volunteers (146,436 women, 72,087 men, 913 unreported) from the Project Implicit website participating between January 1 and December 31, 2012.

Participants were included in this study if: (a) they completed one of the nine relevant IATs, (b) they were over the age of 18, and (c) they were a citizen and resident of the United States. We chose only to use US residents for two reasons: (a) people in the United States are likely to face social pressures to appear egalitarian; it is less clear whether or not people in other countries face such pressures, and (b) individuals from countries other than the United States may be unfamiliar with certain prejudices that are unique to the United States (e.g., Native American stereotypes). Participant ages ranged from 18 years to 89 years (M = 28.9, SD = 11.6). The supplemental materials contain demographic characteristics by individual sample.

Project Implicit Procedure

Visitors typically come to Project Implicit via Internet outlets such as "blogs… personal recommendation, search engines, topically relevant sites that provided a link, [or] as a class or work recommendation or assignment," (Nosek et al., 2007). After consenting to participation, visitors chose from a list of fourteen topics.[1] Participants then completed a demographic questionnaire, an IAT, and measures of their explicit attitudes (all counterbalanced). Upon completion of the study, participants received feedback about their IAT score. They then answered a series of standard questions about their opinions of this feedback. The entire procedure took about 10 minutes.

Measures

Because the data are archival, we did not determine the content of any questions, but used those items we thought best represented the constructs of interest.[2] The supplemental methods contain specific items for each IAT and all items measured and data are available at https://osf.io/y9hiq/wiki/home/.

Defensiveness. Participants indicated the extent to which they agreed with the statements: "Whether I like my IAT score or not, it captures something important about me" (reverse coded), "The IAT reflects something about my automatic thoughts and feelings concerning this topic" (reverse coded), and "The IAT does not reflect anything about my thoughts or feelings unconscious or otherwise;" *Strongly Disagree, Disagree, Agree, Strongly Agree*. We combined these items into an index of feedback derogation for each topic (α's > .70). Examining whether people derogate feedback is an approach to defensiveness taken by many theorists (Sherman, 2013), and these same items have been used to operationalize defensive feedback derogation in at least one other study using data from the Project Implicit website (Howell et al., 2014).

IAT Feedback. The current study incorporated data from both stereotyping and evaluative IATs. Stereotyping IATs (i.e., Black-White/Weapons-Harmless Objects, Gender/Career-Family, Gender/Science-Liberal Arts, Native-White/Foreign-American, Asian-White/Foreign-American) measure the strength of association between the two target categories (e.g., Black and White) and two stereotype categories (e.g., weapons and harmless objects). After completing the study, participants received one of seven types of feedback which indicated that they had a *Slight, Moderate,* or *Strong* implicit association between each category of individuals and the stereotype categories, or that they showed no

[1] We did not examine all of the IATs because either another researcher was using the dataset (N = 1) or the implicit measure was something other than the IAT (e.g., the AMP; N = 4).

[2] We first explored the psychometric properties of these measures in the Black-White/Weapons-Harmless Objects IAT, and then replicated our analyses in a confirmatory manner in all other data sets.

implicit association. Thus, feedback was on a scale ranging from 1 = Strong anti-stereotypical bias ("Your results indicate a strong association between Black and harmless objects and White and weapons"), to 7 = Strong stereotypical bias ("Your results indicate a strong association between White and harmless objects and Black and weapons").

Evaluative IATs (i.e., Gay-Straight/Good-Bad, Young-Old/Good-Bad, Abled-Disabled/Good-Bad, Arab Muslim-Other People/Good-Bad) measure participants' preference between the target categories (e.g., Gay and Straight) by evaluating the association between these categories and words from the evaluative categories good (e.g., joy, pleasure) and bad (e.g., terrible, failure). After completing an evaluative IAT, participants received one of seven types of feedback indicating that they had a *Slight*, *Moderate*, or *Strong* automatic preference for one group over another or that they had no automatic preference. Thus, feedback was on a scale ranging from 1 = Strong pro-minority preference ("Your results indicate a strong implicit preference for gay people compared to straight people"), to 7 = Strong pro-majority preference ("Your results indicate a strong implicit preference for straight people compared to gay people").

Discussion of the content of each individual IAT and specific feedback wording appears in the method sections reported in the supplemental materials. The distribution of participants' feedback on each IAT appears in Table 1.

Explicit Attitudes. Participants completing a stereotyping IAT reported their association between both target groups (e.g., Black and White people) and each stereotype

Table 1. Study 1: Percentage of participants in each category of explicit attitude and implicit feedback for each IAT.

IAT	Attitude	Strong Anti-Stereotype/ Pro-Minority Bias	Moderate Anti-Stereotype/ Pro-Minority Bias	Slight Anti-Stereotype/ Pro-Minority Bias	No Bias	Slight Stereotype/ Pro-Majority Bias	Moderate Stereotype/ Pro-Majority Bias	Strong Stereotype/ Pro-Majority Bias
Black-White/ Weapons-Harmless Objects	Explicit	1.5%	2.4%	3.5%	68.1%	14.8%	6.9%	2.8%
	Implicit	0.7%	3.1%	5.3%	16.2%	17.4%	30.3%	27.1%
Gender/ Career-Family	Explicit	1.5%	2.1%	1.8%	38.9%	26.1%	19.6%	10.1%
	Implicit	0.4%	2.2%	4.2%	15.2%	17.5%	32.9%	27.6%
Gender/ Science-Liberal Arts	Explicit	0.5%	1.5%	2.4%	38.2%	27.3%	22.4%	7.8%
	Implicit	0.9%	3.8%	6.3%	17.6%	17.5%	28.5%	25.3%
Native-White/ Foreign-American	Explicit	16.3%	11.1%	14.2%	42.2%	4.4%	2.9%	2.3%
	Implicit	2.7%	5.3%	9.8%	26.2%	13.5%	20.4%	22.0%
Asian -White/ Foreign-American	Explicit	4.2%	6.7%	11.3%	53.8%	14.4%	6.4%	3.1%
	Implicit	2.5%	6.6%	7.7%	17.3%	15.0%	24.4%	26.5%
Gay-Straight/ Good-Bad	Explicit	1.5%	2.6%	4.9%	55.2%	12.9%	9.1%	10.1%
	Implicit	3.5%	6.7%	7.7%	16.1%	13.8%	22.7%	26.1%
Young-Old/ Good-Bad	Explicit	1.1%	4.1%	10.6%	42.1%	20.2%	13.4%	3.6%
	Implicit	0.0%	2.2%	4.2%	13.9%	15.8%	29.0%	32.4%
Abled-Disabled/ Good-Bad	Explicit	0.1%	1.0%	2.6%%	61.6%	18.8%	9.2%	3.6%
	Implicit	0.1%	2.5%	4.0%	11.6%	12.3%	24.8%	40.6%
Arab Muslim-Other People/ Good-Bad	Explicit	0.4%	1.0%	2.3%	55.7%	22.3%	11.2%	7.1%
	Implicit	5.5%	12.6%	13.7%	26.1%	17.0%	17.2%	8.0%
Average[1]	**Explicit**	**3.0%**	**3.6%**	**6.4%**	**49.3%**	**17.9%**	**11.2%**	**5.6%**
	Implicit	**1.8%**	**5.0%**	**7.0%**	**17.8%**	**15.5%**	**25.6%**	**26.2%**

[1]Average across all 9 IATs, weighted by sample size.

concept (e.g., weapons and harmless objects). For example, participants completing the Black-White/Weapons-Harmless Objects task indicated "how strongly [they] associated weapons (harmless objects) with Black and White people" 1 = *Strongly with White people*, 7 = *Strongly with Black people*. We reverse coded one of the items (here about harmless objects) so that both items mirrored the direction of implicit feedback.

Participants completing an evaluative IAT reported their preference between the target categories on an item that mirrored implicit feedback. For instance, participants completing the Gay-Straight/Good-Bad IAT reported their preference for straight versus gay individuals on a response scale ranging from 1 = *I strongly prefer gay people to straight people* to 7 = *I strongly prefer straight people to gay people*. The distribution of participants' implicit and explicit attitudes appears in Table 1.

Magnitude and Direction of Implicit-Explicit Discrepancy. Consistent with earlier work (Howell et al., 2014), we computed an implicit-explicit discrepancy score by subtracting self-reported attitude (on a seven point scale) from IAT feedback (on a seven point scale) within each study. For IATs with two explicit items (Stereotype IATs), we computed two difference scores for implicit-explicit agreement. We then averaged these scores to create a single measure of discrepancy.

We calculated the *magnitude* of implicit-explicit discrepancy by taking the absolute value of discrepancy. We square-root transformed this value to account for positive skew (Freeman & Tukey, 1950). We also coded the *direction* of explicit-implicit discrepancy as -1 for participants whose implicit bias aligned less with societal bias than their explicit bias (e.g., they were explicitly more favorable toward the majority group than they were implicitly), 0 if implicit and explicit attitudes/stereotypes aligned, and +1 for participants whose implicit bias aligned more with societal bias (i.e., they were implicitly more favorable toward the majority group than they were explicitly). Note that positive direction could have resulted from four different sources (and, conversely, negative direction) including: explicit egalitarianism matched with any implicit societally-consistent implicit bias, explicit societally-inconsistent bias matched with any implicit societally-inconsistent bias, strong societally-inconsistent bias matched with weaker societally-inconsistent bias, and weaker explicit societally-consistent bias matched with stronger implicit societally-consistent bias. The average discrepancy for each IAT appears in Table 2, the distribution of the direction variable for each IAT appears in Table 3. After centering the magnitude variable, we computed an interaction between these two variables, which represented the interaction between magnitude and direction of discrepancy.

Group Membership. For each IAT we also explored group status as a potential moderator. To best detect whether IAT-relevant in-group/out-group status moderated responses, we created a series of contrast codes. That is, we used two contrast codes to compare the reactions of Black (n = 1845) and White (n = 17658) participants in the Black-White/Weapons-Harmless Objects IAT (other = 2391), Native American (n = 808) and White (n = 5887) participants in the Native-White/Foreign-American (other = 1972), Asian (n = 2896) and White (n = 6177) participants in the Asian-White/Foreign-American IAT (other = 2480), and straight (n = 32078) or gay, lesbian, and bisexual (n = 9468) participants in the Gay-Straight/Good-Bad IAT (asexual n = 356). The contrast codes were:

Code 1: *Majority group (e.g., White) = ½, Minority group (e.g., Black) = -½, Other groups (not relevant to the IAT) = 0*

Code 2: *Majority group = 1/3, Minority group = 1/3, Other groups = -2/3.*

Table 2. Study 1: Mean implicit feedback and explicit attitudes, mean magnitude of discrepancy, and correlation between implicit and explicit attitudes among minority group members, majority group members, and in the full sample.

	Minority Group Members				Majority Group Members				Full Sample			
IAT	Implicit M (SD)	Explicit M (SD)	Implicit-Explicit Discrepancy[1]	*r*	Implicit M (SD)	Explicit M (SD)	Implicit-Explicit Discrepancy[1]	*r*	Implicit M (SD)	Explicit M (SD)	Implicit-Explicit Discrepancy[1]	*r*
Black-White/Weapons-Harmless Objects	5.0 (1.5)	4.1 (0.9)	0.9 (1.7)*	.14*	5.5 (1.4)	4.3 (0.7)	1.3 (1.4)*	.16*	5.5 (1.4)	4.2 (0.7)	1.2 (1.5)*	.16*
Gender/Career-Family	5.6 (1.3)	4.8 (0.9)	0.8 (1.4)*	.15*	5.2 (1.4)	4.9 (0.9)	0.4 (1.5)*	.15*	5.5 (1.3)	4.8 (0.9)	0.7 (1.5)*	.14*
Gender/Science-Liberal Arts	5.3 (1.5)	4.9 (0.9)	0.4 (1.5)*	.21*	5.5 (1.4)	4.9 (0.9)	0.6 (1.5)*	.20*	5.3 (1.5)	4.9 (0.9)	0.5 (1.5)*	.21*
Native-White/ Foreign-American	3.4 (1.8)	2.6 (1.6)	0.8 (2.3)*	.11^{+}	4.9 (1.9)	3.4 (1.3)	1.4 (2.1)*	.21*	4.6 (2)	3.3 (1.4)	1.3 (2.1)*	.23*
Asian -White/ Foreign-American	4.3 (1.7)	3.4 (1.3)	0.9 (2.1)*	.10*	5.6 (1.5)	4.4 (1.0)	1.3 (1.6)*	.21*	5.1 (1.7)	4.0 (1.2)	1.2 (1.8)*	.26*
Gay-Straight/ Good-Bad	3.8 (1.8)	3.5 (1.1)	0.3 (1.9)*	.22*	5.5 (1.5)	4.8 (1.1)	0.7 (1.6)*	.31*	5.1 (1.7)	4.5 (1.2)	0.6 (1.7)*	.41*
Young-Old/ Good-Bad	--	--	--	--	--	--	--	--	5.6 (1.3)	4.4 (1.2)	1.3 (1.7)*	.12*
Abled-Disabled/ Good-Bad	5.6 (1.6)	4.2 (1)	1.4 (1.7)*	.13*	5.8 (1.4)	4.5 (0.9)	1.4 (1.5)*	.14*	5.8 (1.4)	4.4 (0.9)	1.4 (1.6)*	.14*
Arab Muslim-Other People/ Good-Bad	--	--	--	--	--	--	--	--	4.2 (1.6)	4.6 (1.0)	-0.4 (1.7)*	.30*
Average[2]	**5.3 (1.4)**	**4.6 (1.0)**	**0.7 (1.6)**	**.17**	**5.5 (1.5)**	**4.6 (1.0)**	**0.9 (1.6)**	**.22**	**5.4 (1.5)**	**4.5 (1.0)**	**0.8 (1.6)**	**.20**

* $p < .001$, $^{+}$ $p < .01$
[1]Significance based on a t-test comparing implicit-explicit discrepancy to 0 (i.e., no discrepancy)
[2]Average across all 9 IATs, weighted by sample size

The purpose of the second code was only to create a full set of contrasts so that the first code examined whether majority and minority groups differed in their reactions to IAT feedback. As such, we did not examine and will not report the results of this secondary contrast (see Judd, McClelland, & Ryan, 2011, for a complete description). We used a single contrast code to compare men (n = 17649; 6825) and women (n = 42065; 16235) in the Gender/Career-Family and Gender/Science-Liberal Arts IATs, and able-bodied participants (n = 13117) and participants with a physical disability (n = 2543) in the Abled-Disabled/Good-Bad IAT. The contrast code was *majority group (e.g., men) = ½, minority group (e.g., women) = -½.* We included a continuous measure of age (M = 28.5 years, SD = 12.0) in the Young-Old/Good-Bad IAT. For the purpose of consistency with the other measures, where larger (and positive) codes were associated with majority group members, we mean-centered and reversed coded age so that younger people received higher values. We did not include group membership analyses for the Arab Muslim-Other People/Good-Bad IAT because that study on Project Implicit did not include demographic variables that allowed for certain identification of Arab Muslim status.

Analyses

For each topic we conducted a single regression where we predicted defensiveness from: (a) the magnitude of the discrepancy between self-reported attitude and IAT feedback, (b) the direction of the discrepancy between self-reported attitude and IAT feedback, (c) participants' group status, and (d) all of the possible interactions between these predictor variables.

Study 1: Results and Discussion

Here we present a report of the overall results of the nine combined studies. Individual results sections for each topical study appear in the supplemental materials. Table 1 shows the distribution of implicit and explicit attitudes across the seven-point scale ranging from strong anti-stereotypical/pro-minority-group bias to strong pro-stereotypical/pro-majority-

Table 3. Study 1: Distribution of direction of implicit-explicit discrepancies among minority group members, majority group members, and in the full sample. Explicit>Implicit indicates that a participant's self-reported attitudes aligned more with societally-consistent bias than did their IAT feedback. Explicit=Implicit indicates that self-reported attitudes and IAT feedback aligned. Explicit<Implicit indicates that a participant's self-reported attitudes aligned less with societally-consistent bias than did their IAT feedback.

	Majority Group Members			Minority Group Members			Entire Sample		
IAT	Explicit> Implicit	Explicit= Implicit	Implicit > Explicit	Explicit> Implicit	Explicit= Implicit	Implicit > Explicit	Explicit> Implicit	Explicit= Implicit	Implicit > Explicit
Black-White/ Weapons-Harmless Objects	21.9%	15.1%	63.0%	13.2%	13.3%	73.5%	14.3%	13.3%	72.4%
Gender/ Career-Family	20.9%	13.9%	65.2%	29.7%	16.9%	53.4%	23.6%	14.8%	61.6%
Gender/ Science-Liberal Arts	29.7%	15.1%	55.2%	25.8%	14.0%	60.2%	28.5%	14.8%	56.7%
Native-White/ Foreign-American	35.0%	12.5%	52.5%	29.3%	8.2%	62.4%	30.8%	9.2%	60.1%
Asian -White/ Foreign-American	24.5%	18.7%	56.9%	13.9%	16.5%	69.6%	17.6%	16.9%	65.5%
Gay-Straight/ Good-Bad	37.5%	20.0%	42.5%	25.9%	22.4%	51.7%	28.9%	21.9%	49.2%
Young-Old/ Good-Bad	--	--	--	--	--	--	20.2%	15.4%	64.4%
Abled-Disabled/ Good-Bad	20.9%	13.3%	65.8%	16.8%	14.7%	68.5%	17.5%	14.5%	68.1%
Arab Muslim-Other People/Good-Bad	--	--	--	--	--	--	45.8%	24.4%	29.8%
Average[1]	**25.0%**	**15.0%**	**59.9%**	**22.5%**	**17.0%**	**60.5%**	**24.0%**	**16.3%**	**59.7%**

[1]Average across all 9 IATs, weighted by sample size.

group bias. The final row presents the sample-size-weighted average distribution of attitudes across IATs. This row reveals that participants' explicit attitudes generally followed a normal-curve-like function, with most people indicating that they are either egalitarian or slightly biased (73.6%). By contrast, implicit attitudes generally followed a negatively-skewed curve, with most people having (and thus learning that they had) moderate or strong stereotypical (e.g., pro-Straight) bias (51.8%), and few learning that they were egalitarian (17.8%).

Table 2 shows average implicit and explicit attitudes/stereotypes, the implicit-explicit correlation, and the average implicit-explicit discrepancy score for minority and majority group members. It also provides the results of a one-sample t-test showing that the average implicit-explicit discrepancy differs from zero. The final row shows the sample-size-weighted average of each of these scores across all IATs. This row reveals that, in general, participants' implicit attitudes/stereotypes aligned more with societal prejudice than did their self-reported attitudes/stereotypes, that implicit and explicit measures generally correlated positively (r = .20, $CI_{95\%}$ = .19-.20), and this correlation was slightly stronger among majority-group members (r = .21, $CI_{95\%}$ = .21-.22) than among minority-group members (r = .17, $CI_{95\%}$ = .16-.18).

Table 3 presents the distribution of the direction variable for minority and majority group members. The final row contains a

sample-size-weighted distribution of the direction variable across all IATs. This row reveals that, across samples, most participants learned that their implicit attitudes/stereotypes aligned more with societal prejudice than they self-reported (59.7%).

Table 4 presents the effect sizes ($r_{partial}$) for each predictor. The final column shows the average sample-size weighted effect size for each predictor across all nine IATs, controlling for all other predictors. To assess whether these effect sizes were different from zero across the entire sample, we examined whether the 95% confidence intervals for these values included zero, which would suggest a non-robust effect. Of the seven predictors, only the confidence interval for the three-way interaction between magnitude of discrepancy, direction of discrepancy, and group status included zero. This finding suggests that each predictor uniquely accounted for meaningful variance in feedback derogation.

Overall, participants displayed a moderate level of defensiveness (M = 2.24 out of 4.0, SD = 0.62). Each of the overall models significantly predicted defensiveness, and omnibus correlations ranged from r = .16 (Asian-White/Foreign-American and Gender/Career-Family) to r = .35 (Gay-Straight/Good-Bad). A sample-size-weighted-average across all IATs yielded an average effect size that was meaningfully larger than zero, r =.18, $CI_{95\%}$ = .18-18. Interestingly, the overall model explained about three times as much variance in defensive response to evaluative IATs r = .28, $CI_{95\%}$ = .27-28, compared to stereotyping IATs, , r = .16, $CI_{95\%}$ =. 15-17.

Regression Results

Magnitude and Direction. As Table 4 shows, participants were generally more defensive to the extent that their self-reported attitude and IAT score differed (i.e., magnitude of discrepancy), $.06 \leq r's_{partial} \leq .19$, and when their implicit bias aligned more with societal prejudice than did their explicit attitudes (i.e., direction of discrepancy), $.03 \leq r's_{partial} \leq .13$.

In all but three samples (i.e., Black-White/Weapons-Harmless Objects; Native-White/Foreign-American IAT, Asian-White/Foreign-American IAT), these main effects were qualified by an interaction between magnitude and direction. An examination of the simple slopes for both

Table 4. Study 1: Results of a regression (partial correlations) predicting defensiveness from magnitude of discrepancy, direction of discrepancy, minority versus majority group status and their interactions across all IATs.

	Black-White/ Weapons-Harmless Objects	Gender/ Career-Family	Gender/ Science-Liberal Arts	Native-White/ Foreign-American	Asian - White/ Foreign-American	Gay-Straight/ Good-Bad	Young-Old/ Good-Bad	Abled-Disabled/ Good-Bad	Arab Muslim-Other People/ Good-Bad	**Average[1, 2] $r_{partial}$ [$CI_{95\%}$]**
Magnitude	.08**	.10**	.08**	-.01	.02	.06**	.10**	.09**	.18**	**.06 [.06, .07]**
Direction	.03**	.03**	.05**	.01	.00	.05**	.06**	.12**	.05**	**.03 [.03, .04]**
Group Status	-.03**	.08**	.04**	.00	-.03*	.08**	.03**	.01	—	**.03 [.03, .04]**
Magnitude x Direction	.01	.01**	.03**	.00	.01	.02*	.04**	.04**	.03*	**.02 [.01, .02]**
Magnitude x Group Status	.01+	.02*	.02*	.03+	.03**	.02**	.00	.00	—	**.01 [.01, .02]**
Direction x Group Status	-.01	-.03**	-.02**	.01	.03*	-.03**	.01	-.01	—	**-.01 [-.02, -.01]**
Magnitude x Direction x Group Status	-.01	-.01*	-.01	.00	.03*	-.02*	.00	.03*	—	.00 [-.01, .00]
Model *R*	**.19**	**.16**	**.19**	**.17**	**.16**	**.35**	**.21**	**.32**	**.19**	**.18 [.18, .18]**

** $p \leq .001$, * $p \leq .01$, $^{+}p < .05$

[1]Average across all 9 IATs, weighted by sample size.

[2] **Bold font** in this column indicates that 95% CI does not include zero

directions of feedback consistently revealed that increased discrepancy was related to increased defensiveness more when participants' feedback indicated that their implicit bias aligned *more* with societal prejudice than did their explicit attitudes, $.08 \leq r's_{partial} \leq .20$, than when their feedback indicated that their implicit bias aligned *less* with societal prejudice than did their explicit attitudes, $.02 \leq r's_{partial} \leq .08$.

Group Status. The role of group status in defensiveness was more nuanced across studies. In four of the eight studies including group status (Gender/Career-Family, Gender/Science-Liberal Arts, Gay-Straight/Good-Bad, Young-Old/Good-Bad), majority group members were more defensive than were minority group members. In two of the studies (Black-White/Weapons-Harmless Objects, Asian-White/Foreign-American) minority group members were more defensive than majority group members. Group status did not affect defensiveness in the Native-White/Foreign-American nor Abled-Disabled/Good-Bad studies.

In seven of the eight studies that included group status, the effects of magnitude and direction were qualified by one or more interaction with group status. Generally speaking, these studies suggested that majority group members were particularly sensitive to feedback discrepancy. Indeed, in each case (excepting the Gender/Career-Family study) the effect of magnitude was strongest among majority group members, $.05 \leq r's_{partial} \leq .17$. In the Asian-White/Foreign American and Gay-Straight/Good-Bad studies, the effect of magnitude was particularly strong among majority group members whose feedback indicated their implicit bias aligned *more* with societal prejudice than did their explicit attitudes, $.09 \leq r's_{partial} \leq .17$.

There were two important exceptions to the evidence that majority group status magnified the effects of magnitude and direction. Both occurred in the studies related to gender stereotypes. First, in the Gender/Career-Family study, the effect of magnitude was particularly strong among women (stigmatized group members) whose feedback indicated their implicit attitudes aligned more with societal prejudice than did their explicit attitudes, $r_{\text{partial}} = .09$. Similarly, in the Gender/Liberal Arts-Sciences study, women, rather than men, were particularly sensitive to feedback that indicated their implicit bias aligned more with societal prejudice than did their explicit attitudes, $r_{\text{partial}} = .07$.

Although our examination of gender was exploratory, it is perhaps not surprising that the gender-related IATs might have elicited different effects of group status than did the other IATs. Indeed, although women are generally the more-stigmatized group in society, believing in gender equality may be particularly important to women; that is, in a way that does not translate to other minority groups. For most IATs, it is likely that majority group members strive for egalitarian beliefs (e.g., gay-straight equality) to appear unbiased, whereas minority group members may feel empowered by minority in-group bias (e.g., gay pride). By contrast, women may be more likely than men to value equality because gender equality, rather than gender pride, is a fundamental tenet of most common approaches to feminism (Beasley, 1999). Nevertheless, future research is necessary to better highlight when and why minority and majority group members might respond differently to IAT feedback.

Study 2

Study 1 suggested that receiving feedback about implicit-explicit discrepancy can produce defensiveness; more specifically, it can incite feedback derogation. Nevertheless, there were two important limitations. First, because the data came from an archival data set, we were only able to examine one specific type of defensiveness—feedback derogation. Although feedback/message derogation is a

common measure of defensiveness (Ruiter, Verplanken, Kok, & Werrij, 2003; Witte & Morrison, 2000), it is possible that the present measure of feedback derogation (i.e., downplaying the personal relevance of the feedback) reflects something beyond simple defensiveness. For example, this measure could have also captured the sentiments of those who simply do not believe that the IAT is valid more generally (e.g., because they do not believe computers can provide meaningful feedback about attitudes, because they think there is a "trick" in the test). This does not explain why derogation would be stronger among those with more implicit-explicit discrepancy; nonetheless, it is an important point to consider. Thus, we designed Study 2 to address this question by including several different defensiveness measures including affective responses, behavioral intentions, and multiple forms of task/feedback derogation.

A second limitation of Study 1 is that it lacked random assignment to topic. Participants were volunteers who personally chose the task in which they participated. An examination of demographic evidence suggests that participants, especially those from minority groups, disproportionally selected self-relevant tasks. For example, although the percentage of Native Americans was about 1-2% across most of the IATs, 9.3% of the participants completing the Native American-White/American-Foreign IAT indicated that they were Native American.

Although defensiveness on self-selected tasks represents an important and ecologically-valid outcome, it is possible that self-selected feedback may lead to either increased or decreased defensiveness compared to randomly assigned feedback. For instance, research suggests that people probably avoid feedback that they think might threaten their self-views or prompt defensiveness (Shepperd & Howell, 2015; Sweeny, Melnyk, Miller, & Shepperd, 2010). So, when participants select the type of IAT they complete, their behavior likely signals curiosity and possible openness to the content. As a result, participants might respond *less* defensively than they would if they were required to receive feedback in a given domain. By contrast, participants may select to receive feedback that they think will confirm their self-views about concepts that are highly central to their self-concept. When they receive feedback that violates this cherished self-view, they may actually respond *more* defensively than those who are randomly assigned to receive that type of feedback. Thus, to more fully examine the role of explicit attitude-implicit feedback discrepancy in defensiveness, we randomly assigned participants to IAT type.

Study 2: Method

Participants

Participants were 1125 adults in the United States participating in exchange for $0.51 on Amazon.com's Mechanical Turk between April 1 and May 1 of 2015. Due to an oversight, 40 people participating on April 1 did not complete any demographic measures (3.5% of the total sample). Among those who completed demographic measures 53.8% were women; 77.1% were White, 9.7% were Black (9.7%), 5.3% were Asian, and 7.9% were something else. Participant ages ranged from 18 to 83 years (M = 36.33, SD = 12.9).

We chose to collect 125 participants per condition (IAT type) for three reasons: (1) it was a sufficient sample size to detect a small (r = .10) relationship between our measure and other forms of defensiveness at .80 power and (2) it gave us sufficient power to detect the average model effect size (R = .18) observed in Study 1 and (3) it was a sufficient meta-analytic sample size (N = 1125) to detect the very small effects of magnitude, direction, and their interaction (r ~ .03) that we observed in Study 1 at .80 power (Cohen, 1988). Because of resource concerns, we chose to power on the effects of magnitude, direction, and their interaction (rs ~ .03) meta-analytically as detecting these effects in each IAT would have cost about 1000% more.

Procedure

After consenting to participate, participants completed a measure of affect and were then randomly assigned to one of the nine IATs from Study 1 (i.e., Black-White/Weapons-Harmless Objects, Gender/Career-Family, Gender/Science-Liberal Arts, Native-White/Foreign-American, Asian-White/Foreign-American, Gay-Straight/Good-Bad, Young-Old/Good-Bad, Abled-Disabled/Good-Bad, or Arab Muslim-Other People/Good-Bad). Participants then completed a measure of their self-reported attitudes, took the IAT, and received IAT feedback that was identical to that which was presented in Study 1. Next, all participants completed the defensiveness measure from Study 1, an additional measure of affect, and several new defensiveness measures (detailed below).

Measures

Original Defensiveness Measure. As in Study 1, participants indicated the extent to which they agreed with the statements: "Whether I like my IAT score or not, it captures something important about me" (reverse coded), "The IAT reflects something about my automatic thoughts and feelings concerning this topic" (reverse coded), and "The IAT does not reflect anything about my thoughts or feelings unconscious or otherwise;" *Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree*. We combined these items into an index of feedback derogation for each topic (α = .82; M = 2.19, SD = 0.99).

IAT Derogation. We included four items that assessed derogation of the IAT task, rather than feedback itself. Specifically, participants indicated whether they believed the IAT "is a valid measure of [their] attitudes" (reverse coded), "doesn't really measure anything important," "is a valid measure of [their] bias" (reverse coded), and "is meaningless" (*1* = Strongly Disagree, *7* = Strongly Agree; α = .84; M = 3.91, SD = 1.31).

Feedback Derogation. Six items assessed participants' derogation of their personal feedback. Specifically participants indicated the extent to which they agreed with six statements beginning with the stem "the implicit attitude feedback I just received…" and ending with (1) "is an accurate reflection of my attitudes" (reverse coded), (2) "is based in scientific evidence" (reverse coded), (3) "does not represent my true values," (4) "is distorted" (5) "is exaggerated," (6) "is too extreme" (*1* = Strongly Disagree, *7* = Strongly Agree; α = .85; M = 3.81, SD = 1.23).

Behavioral Intentions. Consistent with research suggesting that people who respond defensively are less willing to change their behavior as a result of feedback (Sherman, Nelson, & Steele, 2000; Witte, 1992), we included the item "The implicit attitude feedback I just received will affect my behavior" (*1* = Strongly Disagree, *7* = Strongly Agree; M = 2.83, SD = 1.57).

Mood. Consistent with earlier work showing that taking the IAT can influence people's mood states, and that these affective shifts may signal defensiveness (Monteith et al., 2001), we examined people's mood using items from the Positive and Negative Affect Schedule expanded form (PANAS-X; Watson & Clark, 1999) before (Time 1) and after (Time 2) they completed the IAT. Specifically, participants indicated the extent to which they felt positive (i.e., interested, excited, strong, enthusiastic; α_{Time1} = .83, M_{Time1} =2.58, SD_{Time1} = 0.93; α_{Time2} = .86, M_{Time1} =2.38, SD_{Time1} = 1.00) and negative (i.e., distressed, upset, guilty, scared, and hostile; α_{Time1} = .82, M_{Time1} =1.34, SD_{Time1} = 0.58; α_{Time2} = .85, M_{Time2} =1.33, SD_{Time2} = 0.59) emotions "currently" ("*Very slightly or not at all", "A little", "Moderately", "Quite a bit", "Extremely*). To assess emotional reactions to the IAT we focused on affect at Time 2 controlling for affect at Time 1 in all analyses.

Reactive Affect. Consistent with evidence suggesting that the content of IAT feedback, rather than just the task itself, can make people feel bad (Howell et al., 2013), we

examined participants' affective reactions to the feedback they received. Participants indicated the extent to which they agreed with the statements "the implicit attitude feedback I just received made me feel bad" (M = 2.88, SD = 1.67) and "the implicit attitude feedback I just received made me feel happy" (M = 3.60, SD = 1.41; *1* = Strongly Disagree, *7* = Strongly Agree). These two items correlated only moderately $r(1124) = -.36$, $p < .001$, so we treated them separately.

Expectations Met. Participants indicated the extent to which "the implicit attitude feedback [they] received [was] about what [they] expected." (*1* = Strongly Disagree, *7* = Strongly Agree; M = 4.19, SD = 1.62).

IAT Feedback, Explicit Attitudes, and Magnitude and Direction of Implicit-Explicit Discrepancy. All measures related to implicit-explicit discrepancy were identical to Study 1 and appear in the online supplement with one exception. Specifically, in Study 2, all participants completing Stereotyping IATs all indicated their explicit attitudes on a single 1-7 scale ranging from 1 = Strong anti-stereotypical bias (e.g., "strong association between Black and harmless objects and White and weapons"), to 7 = Strong stereotypical bias (e.g., "strong association between White and harmless objects and Black and weapons"). We chose to do so to ensure that implicit feedback directly aligned with explicit reports in Study 2.

Group Membership. Unlike Study 1, there were not enough minority group members (e.g., LGB, disabled, Native American, Asian) in the studies to consistently test the effect of group status. As such we chose to focus only on the effects of magnitude, direction, and their interaction in Study 2.

Study 2: Results

Analyses

We conducted analyses in four steps. First, we examined the convergent validity of the measure used in Study 1. Specifically, we examined the correlations between the original measure and all other defensiveness measures. Second, we examined the effects of magnitude, direction, and their interaction on each form of defensiveness for each IAT type using regression, just as in Study 1. Third, we used multilevel modeling to examine the fixed effects of magnitude, direction, and their interaction (Level 1) controlling for IAT-type variance in the average level of each variable (Level 2). Thus, we estimated a model with fixed effects of magnitude, direction, and their interaction, and a fixed and random effect of the intercept[3]. The final model was:

$$\text{Outcome}_{ij} = \gamma_{00} + \gamma_{10}\,\text{Magnitude}_{ij} + \gamma_{20}\,\text{Direction}_{ij} + \gamma_{30}\,\text{MagnitudexDirection}_{ij} + u_{0j} + e$$

Finally, we used the MEDIATE macro for SPSS (Hayes & Preacher, 2014) to examine the indirect effects of magnitude, direction, and their interaction on each of the outcomes. These analyses allowed us to examine whether the defensive responses to the IAT in Study 1 were simply a proxy for another form of defensiveness, or whether IAT derogation (as measured by the original measure) might have downstream consequences for other defensiveness outcomes. Specifically, as Panel 1 of Figure 1 shows, we first examined whether each of the three predictors (magnitude, direction, magnitude x direction) had an indirect effect on each of the measures added in Study 2 (i.e., IAT derogation, feedback derogation, behavioral intentions, positive mood, negative mood, positive reactive affect, negative reactive affect, and expectations) via the original measure of defensiveness. A significant indirect effect, in this case, would suggest that explicit-feedback discrepancy led to defensiveness (as conceptualized by the original defensiveness measure from Study 1),

[3] We chose not to include random effects of magnitude, direction, and their interaction, because doing so would have restricted our fixed DF to only 9, making it vastly underpowered to detect the expected effects.

Table 5. Study 2: Percentage of participants in each category of explicit attitude and implicit feedback for each IAT

IAT	Attitude	Strong Anti-Stereotype/ Pro-Minority Bias	Moderate Anti-Stereotype/ Pro-Minority Bias	Slight Anti-Stereotype / Pro-Minority Bias	No Bias	Slight Stereotype / Pro-Majority Bias	Moderate Stereotype/ Pro-Majority Bias	Strong Stereotype / Pro-Majority Bias
Black-White/ Weapons-Harmless Objects	Explicit	1.8%	2.7%	2.7%	67.9%	16.1%	6.3%	2.7%
	Implicit	0.9%	3.5%	3.5%	19.5%	17.7%	31.9%	23.0%
Gender/ Career-Family	Explicit	3.1%	1.6%	3.1%	44.5%	18.8%	15.6%	13.3%
	Implicit	0.0%	0.0%	6.2%	17.1%	20.9%	27.1%	28.7%
Gender/ Science-Liberal Arts	Explicit	0.0%	0.8%	0.8%	59.2%	20.8%	12.3%	6.2%
	Implicit	0.0%	3.1%	8.5%	21.5%	14.6%	29.2%	23.1%
Native-White/ Foreign-American	Explicit	18.6%	15.0%	21.2%	29.2%	8.0%	2.7%	5.3%
	Implicit	4.3%	7.8%	7.8%	20.9%	19.1%	21.7%	18.3%
Asian -White/ Foreign-American	Explicit	7.1%	2.7%	7.1%	36.3%	28.3%	13.3%	5.3%
	Implicit	0.9%	0.9%	5.3%	15.9%	14.2%	27.4%	35.4%
Gay-Straight/ Good-Bad	Explicit	1.9%	2.8%	1.9%	50.0%	15.1%	2.8%	25.5%
	Implicit	0.0%	0.9%	2.8%	14.0%	18.7%	29.9%	33.6%
Young-Old/ Good-Bad	Explicit	5.0%	10.8%	15.8%	47.5%	10.0%	5.0%	5.8%
	Implicit	1.7%	0.8%	4.2%	15.0%	20.8%	34.2%	23.3%
Abled-Disabled/ Good-Bad	Explicit	6.1%	0.0%	2.6%	48.2%	14.0%	15.8%	13.2%
	Implicit	0.0%	1.8%	0.0%	6.1%	10.5%	23.7%	57.9%
Arab Muslim-Other People/ Good-Bad	Explicit	0.0%	0.9%	0.0%	52.8%	24.1%	8.3%	13.9%
	Implicit	3.7%	5.5%	8.3%	21.1%	26.6%	23.9%	11.0%
Average[1]	**Explicit**	**4.9%**	**4.0%**	**6.6%**	**47.9%**	**17.4%**	**8.9%**	**10.2%**
	Implicit	**1.2%**	**2.7%**	**5.2%**	**16.9%**	**18.1%**	**27.7%**	**28.2%**

[1]Average across all 9 IATs, weighted by sample size.

which in turn led to the outcomes. We also evaluated the reverse indirect effect (Figure 1, Panel 2). That is, we examined whether magnitude, direction, and their interaction had an indirect effect on the original measure of defensiveness through the measures added in Study 2. A significant indirect effect in this case would mean that explicit-feedback discrepancy led to increases in one of the new outcomes (e.g., a violation of expectations), which led to an increase in defensiveness (as conceptualized by the original defensiveness measure from Study 1). In this final step, we present the results collapsed across IAT types. In each analysis, we used 10,000 bootstrapped samples and obtained an estimate of the effect and a Monte-Carlo-simulation-based 95% confidence interval.

Descriptive Statistics

Table 5 shows the distribution of implicit and explicit attitudes across the seven-point scale ranging from strong anti-stereotypical/pro-minority-group bias to strong pro-stereotypical/pro-majority-group bias. The final row presents the sample-size-weighted average distribution of attitudes across IATs. This row reveals that participants' explicit attitudes generally mirrored those in Study 1, with most people indicating that they are either egalitarian or slightly biased (65.3%). Again, as in Study 1, most people learned that they had moderate or strong stereotypical (e.g., pro-Straight) bias (55.9%), and few learned that they were egalitarian (16.9%).

Table 6 shows average implicit and explicit attitudes/stereotypes, the implicit-explicit correlation, and the average implicit-explicit discrepancy score for minority and majority group members. Consistent with Study 1, in general, participants' implicit

attitudes/stereotypes aligned more with societal prejudice than did their self-reported attitudes/stereotypes and implicit and explicit measures generally correlated positively (r = .16, $CI_{95\%}$ = .10-.21).

Table 7 presents the distribution of the direction variable. The final row shows that, consistent with Study 1, most participants learned that their implicit attitudes/stereotypes aligned more with societal prejudice than they self-reported (61.3%).

Correlation With Other Defensiveness Measures

Table 8 shows the correlation between all of the measures of defensiveness as well as with expectations. The final row shows the average correlation between the each measure and the other measures. As the first column shows, the measure of defensiveness used in Study 1 correlated significantly with all other measures of defensiveness except the two measures assessing negative mood. The highest correlation was with the measure of feedback derogation, $r(1125) = .74$, and the lowest significant correlation was with positive mood $r(1125) = -.18$. Taken together with the non-correlations with negative mood, this suggests that the original measure of defensiveness

Table 6. Study 2: Mean implicit feedback and explicit attitudes, mean magnitude of discrepancy, and correlation between implicit and explicit attitudes.

IAT	Implicit M (SD)	Explicit M (SD)	Implicit- Explicit Discrepancy[1]	*r*
Black-White/Weapons-Harmless Objects	5.4 (1.4)	4.2 (0.9)	1.14 (1.5)**	.24*
Gender/Career-Family	5.6 (1.2)	4.7 (1.4)	0.8 (1.7)**	.18*
Gender/Science-Liberal Arts	5.3 (1.4)	4.6 (1)	0.66 (1.5)**	.29**
Native-White/ Foreign-American	4.8 (1.7)	3.2 (1.6)	1.57 (2.3)**	.04
Asian -White/ Foreign-American	5.7 (1.4)	4.4 (1.4)	1.28 (2.1)**	-.12
Gay-Straight/ Good-Bad	5.7 (1.2)	4.8 (1.5)	0.9 (1.6)**	.28**
Young-Old/ Good-Bad	5.5 (1.3)	3.9 (1.4)	1.63 (1.7)**	.17+
Abled-Disabled/ Good-Bad	6.3 (1.1)	4.6 (1.5)	1.64 (1.7)**	.09
Arab Muslim- Other People/ Good-Bad	4.8 (1.5)	4.8 (1.1)	-0.02 (1.8)	.12
Average[2]	5.4 (1.4)	4.4 (1.4)	1.07 (1.8)**	.16**

** $p < .001$, * $p < .05$, $^{+}$ $p < .10$

[1]Significance based on a t-test comparing implicit-explicit discrepancy to 0 (i.e., no discrepancy)

[2]Average across all 9 IATs, weighted by sample size.

Table 7. Study 2: Distribution of direction of implicit-explicit discrepancies. Explicit>Implicit indicates that a participant's self-reported attitudes aligned more with societally-consistent bias than did their IAT feedback. Explicit=Implicit indicates that self-reported attitudes and IAT feedback aligned. Explicit<Implicit indicates that a participant's self-reported attitudes aligned less with societally-consistent bias than did their IAT feedback.

	Entire Sample		
IAT	Explicit > Implicit	Explicit = Implicit	Implicit > Explicit
Black-White/ Weapons-Harmless Objects	14.3%	16.1%	69.6%
Gender/Career-Family	23.4%	24.2%	52.3%
Gender/Science-Liberal Arts	19.2%	27.7%	53.1%
Native-White/ Foreign-American	19.5%	14.2%	66.4%
Asian -White/ Foreign-American	17.7%	18.6%	63.7%
Gay-Straight/Good-Bad	17.0%	28.3%	54.7%
Young-Old/Good-Bad	10.8%	10.8%	78.3%
Abled-Disabled/Good-Bad	9.6%	16.7%	73.7%
Arab Muslim-Other People/Good-Bad	38.9%	21.3%	39.8%
Average[1]	**18.9%**	**19.8%**	**61.3%**

[1]Average across all 9 IATs, weighted by sample size.

Table 8. Study 2: Correlations between all study outcomes.

		1	2	3	4	5	6	7	8	9
1	Original Measure									
2	IAT Derogation	**.74**								
3	Feedback Derogation	**.67**	**.76**							
4	Behavioral Intentions	**-.28**	**-.34**	**-.18**						
5	Positive Mood [B]	**-.18**	**-.18**	**-.21**	-.01					
6	Negative Mood [B]	-.01	.05	**.14**	**.11**	-.06				
7	Positive Reactive Affect	**-.32**	**-.36**	**-.46**	**.12**	**.28**	**-.28**			
8	Negative Reactive Affect	.02	**.07**	**.28**	**.36**	**-.20**	**.34**	**-.36**		
9	Expectations Met	**-.44**	**-.45**	**-.58**	**.08**	**.15**	**-.17**	**-.27**	**.43**	
	Average [$CI_{95\%}$] [A]	**.33 [.31, .35]**	**.37 [.35, .39]**	**.41 [.39, .43]**	**.18 [.17, .20]**	**.16 [.14, .18]**	**.29 [.27, .31]**	**.31 [.29, .32]**	**.26 [.24, .27]**	**.32 [.30, .34]**

Notes: **bolded** $p \leq .01$; [A] Average correlation with all other measures. Uses absolute values of 4, 5, 7, and 9, as they are expected to correlate negatively with defensiveness. [B] Controlling for mood state prior to the IAT

likely reflected more than just an affective reaction to the news. Overall, it correlated moderately with all of the other measures, on average, $r(1125) = .33$, suggesting that the original measure was convergent valid.

Regression Outcomes

Table 9 shows the associations ($r_{partial}$) between magnitude, direction, and their interaction on each of the defensiveness measures. The final column shows the results of the multilevel model collapsing across IAT condition. As the first rows in the final column show, consistent with predictions, magnitude, $r_{partial} = .12$, direction, $r_{partial} = .06$, and their interaction, $r_{partial} = -.08$, had a significant effect on defensiveness as measured in Study 1. The results suggested that people were more defensive to the extent that their explicit attitudes and implicit attitudes differed (magnitude) and when they received feedback indicating their implicit attitudes aligned more with societal bias than did their explicit attitudes. Interestingly, the interaction effect emerged in a direction opposite to that of the first study. That is, magnitude of discrepancy influenced defensiveness most when people learned that their implicit bias aligned *less* with societal pressure than did their explicit bias. This suggests that feedback indicating that one's implicit bias aligned more with societal pressure than did one's explicit bias produced increased threat-response even when the magnitude of that discrepancy was small. The 95% confidence intervals surrounding the meta-analytic estimates also suggested that the effects were larger in Study 2 than they were in Study 1 ($\Delta r_{partial} > .03$). We return to both of these unanticipated findings in the discussion.

Examining the final column in the remaining rows reveals that magnitude, $rs_{partial} > .06$, direction, $rs_{partial} > .07$, and their interaction, $rs_{partial} > .06$, similarly influenced IAT derogation, feedback derogation, negative mood after the task, and ratings regarding the extent to which the feedback violated their expectations[4]. Greater magnitude of explicit-feedback discrepancy, $rs_{partial} = .12$, and learning that one's implicit attitudes aligned more with societal bias than did one's explicit attitudes, $rs_{partial} > .17$, but not the interaction

Table 9. Study 2: Results of a regression (partial correlations) predicting each of the primary outcomes from magnitude of discrepancy, direction of discrepancy their interaction in across all IATs.

	1	2	3	4	5	6	7	8	9	Meta-Analytic Fixed Effect
Original Measure										
Magnitude	.13	.03	.09	-.04	.14	.29**	.08	-.05	.30**	.12 [.05, .18]**
Direction	.05	.07	-.03	-.08	.10	.15	-.07	.17+	.22*	.06 [.00, .12]*
MxD	-.07	.03	-.13**	.09	-.15	-.25*	-.04	.13	.04	-.08 [-.14, -.01]*
IAT Derogation										
Magnitude	.13	.12	.11	.02	-.05	.41**	.16+	-.03	.24*	.14 [.07, .20]**
Direction	.07	.02	.08	.01	.26**	-.06	-.02	.06	.24*	.09 [.02, .15]**
MxD	-.08	-.07	-.14	.09	-.2*	-.12	.02	.14	.00	-.07 [-.13, -.01]*
Feedback Derogation										
Magnitude	.38**	.25**	.17+	.09	.2*	.39**	.19*	-.04	.39**	.23 [.18, .29]**
Direction	.11	.04	.10	-.02	.18+	.18+	.07	.27**	.26**	.14 [.08, .20]**
MxD	-.16+	-.10	-.14	.11	-.16+	-.10	-.08	.15	.09	-.09 [-.15, -.02]**
Behavioral Intentions										
Magnitude	.02	-.01	.14	.16+	.17+	-.08	.17+	.00	-.04	.05 [-.01, .11]
Direction	.02	.11	.07	.05	-.26**	.03	-.06	.03	.04	-.01 [-.07, .05]
MxD	.22*	-.05	.11	-.10	.12	-.05	-.21*	-.11	.00	.00 [-.06, .06]
Positive Mood [AB]										
Magnitude	.13	-.08	-.02	-.04	-.12	-.27**	.02	.08	-.06	-.03 [-.09, .03]
Direction	-.12	.00	-.06	-.12	-.09	-.12	.03	-.15	-.18+	-.08 [-.14, -.02]*
MxD	-.03	.03	-.05	.17+	.00	-.1	.00	.01	.02	.03 [-.03, .09]
Negative Mood [A]										
Magnitude	.18+	.23*	-.04	-.01	-.13	.15	-.02	-.04	.06	.06 [.00, .12]+
Direction	.14	.07	.00	.11	.11	.06	-.03	.15	.25*	.11 [.05, .17]**
MxD	.00	-.13	-.34**	.15	.03	.07	.01	-.03	-.07	-.06 [-.12, .00]+
Positive Reactive Affect										
Magnitude	-.26**	-.26**	-.13	-.10	.02	-.37**	-.10	.15	-.08	-.12 [-.18, -.06]**
Direction	-.28**	-.19*	-.15+	.03	-.33**	-.07	.01	-.19+	-.29**	-.17 [-.23, -.11]**
MxD	-.04	-.07	.06	-.11	.12	.03	.12	-.08	.03	.04 [-.02, .10]
Negative Reactive Affect										
Magnitude	.26**	.19*	.23**	.17+	.09	.07	.23*	-.11	.02	.12 [.06, .18]**
Direction	.15	.12	.21*	.17+	.04	.24*	.05	.29**	.39**	.19 [.13, .25]**
MxD	.17+	-.08	.09	.05	.06	.05	.12	.01	-.04	.03 [-.03, .09]
Expectations Met										
Magnitude	-.44**	-.27**	-.33**	-.09	-.17+	-.54**	-.17+	-.10	-.30**	-.28 [-.34, -.23]**
Direction	.02	.02	.01	.03	-.18+	-.05	-.04	-.21*	-.15	-.07 [-.13, -.01]*
MxD	.05	.10	-.05	-.03	.04	.12	.09	-.13	.03	.06 [.00, .12]*

1. Black-White/ Weapons-Harmless Objects 2. Gender/ Career-Family 3.Gender/ Science-Liberal Arts 4. Native-White/ Foreign-American 5. Asian -White/ Foreign-American 6. Gay-Straight/ Good-Bad 7. Young-Old/ Good-Bad 8.Abled-Disabled/ Good-Bad 9. Arab Muslim- Other People/ Good-Bad.

Notes: [A] Controlling for mood state prior to the IAT [B] Model would not converge with random intercept included, random intercept is removed.

[4] We intentionally did not have enough power to detect the expected effects in the individual

Table 10. Study 2: Indirect effects of magnitude, direction, and their interaction on each of the defensiveness outcomes.

	Indirect effects on all defensiveness measures via the original measure of defensiveness *b* [$CI_{95\%}$]			Indirect effects [$CI_{95\%}$] on the original measure of defensiveness via the other defensiveness measures *b* [$CI_{95\%}$]		
	Magnitude	Direction	Magnitude x Direction	Magnitude	Direction	Magnitude x Direction
IAT Derogation	**.44 [.25, .63]**	**.08 [.01, .16]**	**-.35 [-.62, -.08]**	**.26 [.14, .38]**	**.09 [.03, .14]**	**-.24 [-.44, -.04]**
Feedback Derogation	**.37 [.20, .53]**	**.07 [.004, .14]**	**-.29 [-.51, -.06]**	**.42 [.31, .53]**	**.13 [.08, .19]**	**-.27 [-.46, -.08]**
Behavioral Intentions	**-.22 [-.34, -.12]**	**-.04 [-.08, -.003]**	**.17 [.04, .32]**	-.04 [-.09, .01]	-.004 [-.03, .02]	-.001 [-.09, .09]
Positive Mood	**-.05 [-.09, -.03]**	**-.01 [-.02, -.001]**	**.04 [.01, .08]**	.01 [-.01, .05]	.02 [.004, .04]	-.03 [-.08, .02]
Negative Mood	**-.18 [-.27, -.10]**	**-.03 [-.07, -.002]**	**.14 [.03, .26]**	-.004 [-.02, .01]	-.004 [-.01, .01]	.01 [-.01, .03]
Positive Reactive Affect	**-.18 [-.27, -.10]**	**-.03 [-.07, -.002]**	**.14 [.03, .26]**	**.09 [.04, .14]**	**.07 [.04, .10]**	-.06 [-.15, .02]
Negative Reactive Affect	-.01 [-.06, .04]	-.001 [-.01, .01]	.01 [-.03, .05]	-.003 [-.02, .02]	-.003 [-.02, .01]	.0003 [-.01, .01]
Expectations Met	**-.30 [-.44, -.17]**	**-.06 [-.11, -.003]**	**.23 [.06, .42]**	**.33 [.25, .41]**	**.04 [.01, .08]**	**-.12 [-.24, -.002]**

Notes: **bolded** indicates $p \leq .01$

between the two, were associated with increased negative affect and decreased positive affect in response to the task. Moreover, learning that one's implicit attitudes aligned more with societal bias than did one's explicit attitudes was solely related to decreased positive mood after the task, $r_{partial} = .08$. Implicit-explicit discrepancy did not predict behavioral intentions, $rs_{partial} < .06$.

Indirect Effects

Table 10 shows both sets of indirect effects. The results suggested that there were bi-directional indirect effects of magnitude, *b*s > .08, direction, *b*s > .03, and their interaction, *b*s > .12, on the original measure of defensiveness and IAT derogation, feedback derogation, positive reactive affect, and expectation violations. These results suggest that increased discrepancy between feedback and explicit attitudes, particularly when the feedback suggests that one's attitudes align more with societal bias, simultaneously increased these forms of defensiveness. There were uni-directional indirect effects of magnitude, *b*s > .05, direction, *b*s > .01, and their interaction, *b*s > .04, on behavioral intentions, positive mood, and negative mood via the original measure of defensiveness. That is, increased discrepancy between feedback and explicit attitudes, feedback suggesting that one's attitudes align more with societal bias, and their interaction increased defensiveness, which in turn led to reduced intentions to change one's behavior, reduced positive mood, and increased negative mood.

Is discrepancy necessary?

In Studies 1 and 2, we chose to focus on discrepancy between IAT feedback and self-reported (explicit) attitudes. We expected that receiving feedback that contradicts the way people see themselves prompts them to react defensively. Nevertheless, it is possible that all of these effects are driven by either IAT feedback or self-reported attitudes, but not the discrepancy between the two. Such would be the case if, for example, people who endorse egalitarian attitudes are generally more defensive. Moreover, it could be that receiving feedback that one is implicitly biased, regardless of one's explicit attitudes, prompts defensiveness. To rule out this possible alternative explanation we examined the main effects of self-reported attitudes, IAT feedback, and their interaction on each of the defensiveness-related measures.

Table 11. Results of a regression (partial correlations) predicting each of the primary outcomes from implicit explicit attitudes, and their interaction in across all IATs.

	1	2	3	4	5	6	7	8	9	Meta-Analytic Effect [B,C]
Study 1: Original Measure										
Implicit	.08**	-.01*	.05**	.19**	.03**	.22**	.09**	.23**	.00	**.09 [.03, .16]**
Explicit	-.15**	-.16**	-.14**	.00	-.08**	-.14**	-.16**	-.19**	-.02	**-.12 [-.18, -.06]**
IxE	-.11**	-.10**	-.08**	-.06**	-.09**	-.21**	-.12**	-.11**	-.19**	**-.12 [-.18, -.06]**
Study 2: Original Measure										
Implicit	-.03	.02	.03	.00	.13	.08	-.07	.14	.18+	.05 [-.01, .11]
Explicit	-.23*	-.12	.09+	.09	.05	-.16+	.02	-.11	-.28**	**-.07 [-.13, -.01]**
IxE	-.11	-.03	-.08**	-.04	-.07	-.35**	-.09	.01	-.37**	**-.12 [-.18, -.06]**
IAT Derogation										
Implicit	.02	.07	.15	.09	.08	.15	.00	.13	.18+	**.10 [.03, .16]**
Explicit	-.21*	-.05	.06+	.04	-.05	-.06	-.14	-.01	-.26**	**-.07 [-.13, -.01]**
IxE	-.05	-.06	-.11**	-.07	.17+	-.30**	-.10	-.04	-.32**	**-.10 [-.16, -.04]**
Feedback Derogation										
Implicit	.14	.13	.13+	.14	.13	.33**	.11	.22*	.18+	**.17 [.11, .23]**
Explicit	-.28**	-.11	-.09**	.06	-.09	-.20*	-.06	-.17+	-.26**	**-.13 [-.19, -.07]**
IxE	-.16**	-.23*	-.11**	-.08	-.05	-.31**	-.09	-.01	-.19+	**-.14 [-.20, -.08]**
Behavioral Intentions										
Implicit	.03	.07	.09	-.06	-.01	-.10	.00	-.09	-.04	-.01 [-.07, .05]
Explicit	-.19+	-.04	-.22*	-.13	.09	-.02	.04	-.04	-.12	**-.07 [-.13, -.01]**
IxE	-.01	-.05	-.01	-.14	-.21*	.06	.11	.04	.00	-.02 [-.09, .04]
Positive Mood [A]										
Implicit	-.03	-.07	-.23**	.07	-.19*	-.27**	.09	-.01	-.21*	**-.09 [-.16, -.03]**
Explicit	.05	-.01	-.24**	.06	.03	.18+	.01	.05	-.03	.01 [-.06, .07]
IxE	-.06	.08	.07	.03	-.02	.21*	.09	.02	.03	.05 [-.01, .11]
Negative Mood [A]										
Implicit	.20*	.23**	-.07	.17+	.10	.16	.16+	.09	.20*	**.14 [.08, .20]**
Explicit	-.22*	-.05	.11	-.11	.01	-.08	.16+	-.02	-.03	-.02 [-.08, .04]
IxE	.05	-.06	-.26**	.06	.21*	-.17+	.09	-.03	-.01	-.02 [-.08, .05]
Positive Reactive Affect										
Implicit	-.37**	-.37**	-.2*	-.13	-.31**	-.31**	-.07	-.20*	-.34**	**-.26 [-.32, -.21]**
Explicit	.27**	.18*	.08	.00	-.04	.15	-.19*	.03	-.06	.05 [-.01, .11]
IxE	.01	.19*	.08	.02	-.08	.18+	.16+	-.11	-.03	.05 [-.01, .11]
Negative Reactive Affect										
Implicit	.27**	.14	.27**	.18+	.20*	.27**	.07	.26**	.34**	**.23 [.17, .28]**
Explicit	-.16	-.16+	-.27**	-.16+	.04	-.14	-.02	-.05	-.16+	**-.12 [-.18, -.06]**
IxE	-.23*	-.15+	-.01	-.12	-.10	-.07	-.11	.04	-.07	**-.09 [-.15, -.03]**

1. Black-White/ Weapons-Harmless Objects 2. Gender/ Career-Family 3.Gender/ Science-Liberal Arts 4. Native-White/ Foreign-American 5. Asian -White/ Foreign-American 6. Gay-Straight/ Good-Bad 7. Young-Old/ Good-Bad 8.Abled-Disabled/ Good-Bad 9. Arab Muslim- Other People/ Good-Bad. Notes: [A] Controlling for mood state prior to the IAT [B] Average across all 9 IATs, weighted by sample size. [C] **Bold font** in this column indicates that 95% CI does not include zero.

Table 11 shows these results for each IAT, as well as a meta-analytic effect across IATs (final column). An evaluation of these meta-analytic effects shows that all three

measures of feedback derogation, IAT derogation, and negative reactive affect were all driven by an interaction (i.e., a discrepancy) between self-reported attitudes and IAT feedback. The two mood variables and positive reactive affect were driven primarily by implicit feedback, suggesting that receiving bad news (i.e., evidence that one is not egalitarian) generally influenced these variables. By contrast, intentions to change one's behavior as a result of feedback resulted primarily from explicit attitudes, suggesting that those who endorsed societally non-stereotypic attitudes were primarily those who wanted to change their behavior based on feedback. Notably, the variables that were unaffected by an implicit-explicit interaction were generally uninfluenced directly by aspects of discrepancy, but were often predicted by an indirect effect of discrepancy via feedback derogation. Taken together, this suggests that implicit-explicit discrepancy played a critical role in determining more-prototypical defensiveness operationalizations, but that examining this discrepancy may be less useful than attitude main effects in directly predicting mood, positive reactive affect, and behavioral intentions.

Reactive Affect and Defensiveness: Two opposite paths to behavior?

The findings of Studies 1 and 2 clearly demonstrate that defensiveness is predicted by implicit-explicit attitude discrepancy, particularly feedback suggesting people harbor more societally consistent bias implicitly than they do explicitly. Study 2 also suggested that this defensive responding could, in turn, dampen intentions to engage in behavior that should reduce personal bias. An interesting and unexpected correlation emerged with respect to negative reactive affect. Specifically, although implicit-explicit discrepancy predicted more negative reactive affect, people with greater negative reactive affect had increased perceptions that their feedback would influence their future behavior, r (1125) = .36, $p < .001$. This correlation suggests a positive, rather than a negative, indirect path from unexpected and undesired feedback to behavior through negative reactive affect.

Considering the experience of the participant suggests a possible reason for this correlation. Imagine three participants who take the Gay-Straight/Good-Bad IAT, all of whom believe they are egalitarian and cherish that self-view. Person A receives feedback that she is egalitarian—equally preferring gay and straight people. Person B and Person C both receive feedback that they strongly prefer Straight people to Gay people. Person B responds defensively—derogating the IAT. As such, Person B is thus able to maintain a view of himself as egalitarian. Persons A and C both accept their feedback as accurate. As a result, Person A should neither derogate the IAT nor experience any negative affect, and will maintain a view of herself as egalitarian. Person C, by contrast, accepts her feedback—a challenge to her self-view—and should therefore feel guilty and sad leading her to endorse the item suggesting that "my feedback made me feel bad." Moreover, because Person C wants to be egalitarian, but the IAT suggests she is not, Person C must also be willing to change her behavior to become more egalitarian.

In sum, it is likely that there are two opposing paths from feedback to behavior: one through defensiveness—which suppresses desire to change behavior, and another through negative reactive affect—which promotes desire to change behavior. This hypothesis is consistent with previous work outside the IAT domain (e.g., Devine, Monteith, Zuwerink, & Elliot, 1991; Voils, Ashburn-Nardo, & Monteith, 2002) which shows that experiencing or detecting discrepancies between one's egalitarianism and one's potentially prejudiced behaviors induces feelings of discomfort and, for those low in self-reported prejudiced, guilt. Thus, in the current research, feedback indicating bias should induce greater negative

affect. And based on findings that negative affect (e.g., guilt) can promote self-regulatory attempts—including attempts to reduce prejudice (Amodio, Devine, & Harmon-Jones, 2007)—we expect that negative reactive affect would predict greater willingness to change one's behavior. To explore this possibility, we conducted a simultaneous indirect path model in which Magnitude, Direction, and their Interaction predicted behavioral intentions, through negative reactive affect and our original measure of defensiveness, including data from all of the IATs in Study 2.

Table 12 shows the indirect effects of magnitude, direction, and their interaction on behavioral intentions through reactive affect and defensiveness. Consistent with the hypothesis that there are two opposing paths to behavioral intentions, for both magnitude and direction there was a positive indirect path to behavioral intentions through negative reactive affect. By contrast, there was a negative indirect path of magnitude and direction on behavioral intentions through defensiveness. That is, to the extent that participants' implicit feedback differed from their self-reports, and when they were told they held societally-consistent bias (e.g., they preferred Straight to Gay more implicitly than explicitly), they were both more likely to feel bad and to respond defensively. To the extent that they felt bad, they were more likely to report that the feedback would influence their behavior. To the extent that they responded defensively, they were less likely to report that feedback would influence their behavior. The indirect path from the interaction between magnitude and direction to behavioral intentions was only mediated by defensiveness.

Table 12. Study 2*:* Indirect effects of magnitude, direction, and their interaction on behavioral intentions through defensiveness and negative reactive affect across all IATs.

	Defensiveness	Neg. Reactive Affect
Magnitude	**-.15 [-.23, -.06]**	**.16 [.07, .25]**
Direction	**-.04 [-.08, -.002]**	**.14 [.09, .19]**
Magnitude x Direction	**.17 [.04, .31]**	-.02 [-.15, .13]

Note: **Bold font** indicates that bootstrapped bias corrected 95% CI does not include zero.

Study 3

Studies 1 and 2 both suggested that implicit-explicit attitude discrepancy prompts defensive responding to IAT feedback. Moreover, Study 2 suggested that learning that one is biased will prompt both defensiveness and negative reactive affect, but that these two outcomes might have very different consequences for behavioral intentions. Nevertheless, because both are correlational, they cannot establish the causal role of receiving specific types of feedback. Moreover, our measure of behavioral intentions in Study 2 was weak, at best. Indeed, it simply asked people to indicate whether their feedback would affect their behavior broadly, rather than asking about intentions or willingness to reduce their implicit bias specifically. In Study 3 we aimed to address these concerns directly by assigning people to either receive feedback indicating egalitarianism or feedback indicating they had implicit pro-White bias. Moreover, we examined participants' willingness to engage in a variety of specific behaviors described as reducing implicit pro-White bias.

Participants

Participants were 413 White adults in the United States participating in exchange for $1.01 on Amazon.com's Mechanical Turk between May 1 and July 1 of 2016. Participants ranged in age from 20 to 71 (M = 37.07, SD = 10.96). Due to an error on Mechanical Turk, 453 adults initially took the study, but 38 of those were the same people taking the study twice. We identified these duplicate participants and removed their second response from the data.

Procedure

After consenting to participate, participants completed the Black-White/Good-Bad IAT. They were then randomly assigned to one of two conditions. In the *Egalitarian* condition they read that they had "no automatic preference for White relative to Black people." In the *Bias* condition they read that they had a

"strong automatic preference for White relative to Black people." The feedback was worded and formatted to mirror that in the correlational studies. Next, all participants completed measures of defensiveness, reactive affect, and behavioral intentions.

Measures

IAT Derogation. We included four items that assessed derogation of the IAT task, rather than feedback itself. Specifically participants indicated whether they believed the IAT "is a valid measure of [their] attitudes" (reverse coded), "doesn't really measure anything important," "is a valid measure of [their] bias" (reverse coded), and "is meaningless" (*1* = Strongly Disagree, *7* = Strongly Agree; α = .88; *M* = 3.98, *SD* = 1.61).

Feedback Derogation. Six items assessed participants' derogation of their personal feedback. Specifically participants indicated the extent to which they agreed with six statements beginning with the stem "the implicit attitude feedback I just received…" and ending with (1) "is an accurate reflection of my attitudes" (reverse coded), (2) "is based in scientific evidence" (reverse coded), (3) "does not represent my true values," (4) "is distorted" (5) "is exaggerated," (6) "is too extreme" (*1* = Strongly Disagree, *7* = Strongly Agree; α = .92; *M* = 3.85, *SD* = 1.76).

Reactive Affect. Consistent with evidence suggesting that the content of IAT feedback, rather than just the task itself, can make people feel bad (Howell et al., 2013), we examined participants' affective reactions to the feedback they received. Participants indicated the extent to which they agreed with the statements "the implicit attitude feedback I just received made me feel bad" (*M* = 2.78, *SD* = 2.01) and "the implicit attitude feedback I just received made me feel good" (*M* = 4.13, *SD* = 2.12; *1* = Strongly Disagree, *7* = Strongly Agree). These two items were highly correlated, $r(409) = -.58$, $p < .001$, so we combined them into a single index of negative reactive affect (*M* = 3.46, SD = 1.84).

General Behavioral Intentions. As in Study 2, participants indicated the extent to which they agreed that their IAT feedback would "affect [their] behavior."

Desire to Change. Participants indicated the extent to which they were "eager to learn how to change [their] implicit bias," "not interested in learning how to change [their] implicit bias" (reverse coded) and had "a responsibility to change [their] implicit bias" (1 = Strongly Disagree, *7* = Strongly Agree; α = .89; *M* = 2.83, *SD* = 1.57).

Behavioral Willingness. Participants indicated the extent to which they were willing to engage in 13 behaviors that "researchers have proposed… for reducing implicit racial bias" (*1* = Very Unwilling, *7* = Very Willing). Example items included "make an active effort to reverse those thoughts when a negative stereotype about Black people comes to mind," "acknowledge that implicit bias can influence my judgments" and "seek out situations to have personal contact with Black people" (α = .93; *M* = 5.24, *SD* = 1.61).

Hypotheses and Analysis

We predicted that people who learned that they were biased would be more likely both to feel bad and to respond defensively. As such, we conducted an independent-samples *t*-test comparing reactive affect and defensiveness between the two feedback groups. We also expected defensiveness and reactive affect to predict behavioral willingness. Specifically, we expected that people who responded the most defensively to the bias condition would be least willing to engage in bias-reducing behaviors. By contrast, we expected that people who felt the worst in response to the bias condition would be the most willing to engage in bias-reducing behaviors. We tested these two hypotheses by examining the simultaneous indirect effects of feedback condition on behavioral intentions and willingness through defensiveness and

reactive affect[5].

Results

Main Effect of Feedback Type. Table 13 shows the bivariate correlations between all of our variables. As expected, participants in the bias condition derogated the IAT (M = 4.66, SD = 1.50), derogated their feedback (M = 5.11, SD = 1.39), and felt worse (M = 4.81, SD = 1.40) than did those in the egalitarian condition (IAT derogation: M = 3.36, SD = 1.45; feedback derogation: M = 2.67, SD = 1.16; negative reactive affect: M = *2.18*, SD = 1.17), *t's*(410) >8.97, *p's* < .001, *d's* > 0.89. Consistent with the notion that those receiving biased feedback would be motivated to do something, whereas those who received egalitarian feedback would not, participants in the bias feedback condition generally indicated that the information would change their behavior (M = 3.27, SD = 1.96) and had a greater desire to change their implicit bias (M = 4.23, SD = 1.87) than did those in the egalitarian feedback condition (general behavioral intention: M = 2.86, SD = 1.47; desire to change: M = 3.73, SD = 1.74), *t's*(410) > -2.22, *p's* < .03, *d's* > 0.24. Participants' willingness to engage in implicit-bias reduction behaviors did not differ between conditions (bias feedback: M = 5.18, SD = 1.34; egalitarian feedback: M = 5.28, SD = 1.27), $t(410) = 0.81$, $p = .42$, $d = 0.08$.

Indirect effects of condition on behavioral willingness and intentions. As Table 13 shows, IAT derogation and feedback derogation were highly correlated, $r = .73$. As such, we treated them as a single index of defensiveness for examining the indirect effects. Examining them separately does not change the pattern of results reported here. Table 14 shows the indirect effect of condition on general behavioral intentions, desire to change one's implicit attitudes, and willingness to engage in implicit-prejudice reducing behaviors. The pattern of results was consistent across all three outcomes.

There was a negative indirect effect of condition on intentions through defensiveness: those who responded defensively to the bias condition were unlikely to intend to change their behavior, were unlikely to desire to change their attitudes, and were the most unwilling to engage in implicit-prejudice reducing behaviors. By contrast, there was a positive indirect effect of condition on intentions through negative reactive affect: Those who felt the worst in response to the bias condition were the most likely to say that the feedback would influence their behavior, desired the most to change their implicit bias, and were the most willing to engage in prejudice-reduction strategies. None of the reverse paths (from condition through

[5] We originally pre-registered predictions for this study at *link to follow publication.* In that document, we predicted that condition would have a negative direct effect on behavioral intentions. That is, we predicted that those who learned they were biased would be less willing to change their behaviors. After collecting slightly more than half of the data, we examined the results to decide whether to continue to pay to collect data and realized two things. First, the direct effect of condition on intentions was not significant, but the indirect effect was significant. Unfortunately, despite having already written the results for Study 2 (which included an indirect effect), and having discussed this path in emails, we failed to include an indirect path from condition to behavior in our initial preregistration. Second, in conversations with colleagues about how an indirect effect could be significant without a direct effect (such a finding typically indicates an opponent process) we decided to propose an affective route to increased behavioral intention. At that time, we first investigated the path from discrepancy to behavior through reactive affect in the Study 2 data. Then, after we had collected all data, we confirmed both expected indirect routes in the Study 3 data.

Table 13. Study 3: Correlations between all outcome measures as well as condition.

	1	2	3	4	5	6
1. Condition[a]	-					
2. IAT Derogation	**.41**	-				
3. Feedback Derogation	**.69**	**.73**	-			
4. Neg. Reactive Affect	**.72**	**.43**	**.68**	-		
5. General Intentions	**.11**	**-.30**	-.08	**.23**	-	
6. Desire to Change	**.14**	**-.24**	.02	**.30**	**.60**	-
7. Behavioral Willingness	-.04	**-.29**	**-.13**	.08	**.35**	**.52**

Note: [a]Condition is coded -.5= egalitarian feedback, .5 = feedback indicating a strong preference for White over Black. **Bold font** indicates correlation is significant at $p < .001$.

Table 14. Study 3: Indirect effects of condition on general behavioral intentions, desire to change, and behavioral willingness through defensiveness and negative reactive affect.

	Defensiveness	Neg. Reactive Affect
General Intentions	**-1.20 [-1.54, -0.90]**	**1.29 [.96, 1.68]**
Desire to Change	**-0.99 [-1.33, -0.71]**	**1.50 [1.14, 1.86]**
Willingness	**-0.62 [-0.84, -0.42]**	**0.67 [0.38, 0.98]**

Note: **Bold font** indicates that bootstrapped bias corrected 95% CI does not include zero.

intentions/willingness) to defensiveness or reactive affect were significant, suggesting a unidirectional process.

Moderation by actual IAT score. Given that people have some insight into their implicit attitudes (Hahn, Judd, Hirsh, & Blair, 2014) and appear to detect how they are performing on the IAT (Monteith et al., 2001), it is possible that participants low in bias who learned they were high in bias (or vice versa) were particularly likely to derogate their IAT feedback—as it was actually untrue. Moreover, it is possible that merely performing poorly on the IAT (high true IAT scores), regardless of any actual feedback elicited all of the effects we observed, consistent with Monteith et al. (2001). To investigate this alternative explanation, we examined the moderating role of actual IAT scores on the relationship between condition and defensiveness. Neither actual IAT scores, $b = 0.41$, SE = .26, $t = 1.56$, $p = .12$, nor the interaction between IAT scores and condition, $b = -0.59$, SE = .37, $t = -1.61$, $p = .11$, were significantly related to defensiveness. These findings suggest that feedback condition, $b = 2.18$, SE = .23, $t = 9.52$, $p < .001$, rather than actual IAT performance was most responsible for the defensive responses we observed.

Discussion

This research provides the first investigation of defensive responses to direct feedback about one's level of implicit bias. Results from three studies suggest that people respond defensively to evidence that contradicts their self-views, particularly if it indicates that they are not egalitarian. The first two large-scale studies, demonstrate that greater discrepancy between self-reported attitudes and implicit attitudes prompts defensive reactions, in the form of derogating feedback, derogating the IAT, and negative emotional reactions. Participants were more defensive to the extent that their implicit and explicit attitudes were discrepant, and were especially defensive when their feedback indicated they were implicitly less egalitarian than they indicated explicitly. The results of the first study also showed that this effect was, with only a few exceptions, strongest among: (a) those who received feedback indicating their implicit attitudes align more with societal prejudice than do their explicit attitudes, and (b) members of majority groups. The third study shows how learning that one is implicitly biased causally produces IAT derogation, feedback derogation, and negative reactive affect. These results are consistent with the idea that people want their self-views to be verified and that they want to see themselves in a

positive light (Sedikides & Gregg, 2008; Swann, Rentfrow, & Guinn, 2003).

The second and third studies also tested whether defensive responses decrease the likelihood of prejudice reduction, investigating people's intentions to change their behavior as a result of IAT feedback. They showed that people who respond more defensively to IAT feedback are less likely to intend to change their behavior, report less desire to influence their implicit bias, and are most unwilling to engage in behaviors that could reduce their implicit bias. By contrast, when participants experienced negative reactive affect (i.e., when their IAT feedback made them feel bad), they were more likely to want to change their behavior. This last finding suggests that learning that one is implicitly biased can both increase and decrease egalitarian behavior, depending on whether one accepts the IAT feedback as valid.

Three unexpected findings emerged. First, the effect sizes in the second study were notably larger than they were in the first study. One plausible explanation for this surprising result is that participants in the second study did not self-select a topic about which to receive feedback. Participants in the first study intentionally accessed the Project Implicit website and chose the task that they wanted to complete. Thus, they may have been (at least somewhat) open to the idea that they could receive undesirable feedback about their implicit attitudes. In fact, the Project Implicit site explicitly warns users of the possibility of undesired feedback. By contrast, participants in the second study did not know that they were going to receive feedback upon initially accessing the study (prior to the consent form) and were randomly assigned to the IAT they completed. As a result, implicit-explicit discrepancy may have impacted participants in the second study more, leading to increased effect sizes.

The second surprising finding was that, in Study 2, the interaction between magnitude and direction reversed in valence. In both studies, the simple main effects of magnitude remained positive supporting initial hypotheses: higher magnitude and greater implicit than explicit bias predicted greater defensiveness. However, in Study 1, magnitude influenced defensiveness most for those who read that their implicit attitudes aligned *more* with societal attitudes than did their self-reported attitudes. But in Study 2, magnitude influenced defensiveness most for those who read that their implicit attitudes aligned *less* with societal attitudes than did their self-reported attitudes. Like the observed effect sizes, this interaction may have resulted from the random assignment to task introduced in Study 2.

In Study 2, people did not select the domain in which they received feedback, and were not necessarily prepared to receive IAT feedback in the domain that they did. As such, it could be that the effect of direction overwhelmed the effect of magnitude to some extent. That is, simply learning that their implicit attitudes align more with societally-consistent bias than their explicit attitudes might have been enough to prompt increased defensiveness, regardless of magnitude. By contrast, people were more sensitive to the magnitude of self-view disconfirmation when they learned that their implicit attitudes aligned *less* with societally-consistent bias than their explicit attitudes.. Still, future research is necessary to better understand the interactions between magnitude and direction.

The final surprising finding involved inconsistencies across the effects involving group in Study 1. Indeed, although the meta-analytic effects we observed are well-powered, the findings within each IAT are quite inconsistent for group status. Although we think this is, again, a problem of self-selection of IATs (e.g., minority groups disproportionally selecting self-relevant tasks; people who might be particularly threatened avoiding certain tasks), we do not have

sufficient data to address this explanation. As such, we encourage caution when interpreting our group status results and urge researchers interested specifically in the role of group status to replicate these effects. New inquiries should both recruit sizable populations of IAT-relevant groups (e.g., minority oversampling) and randomly assign them to take the IATs.

Implications and Applications

The present research reveals IAT feedback as a potential intervention to reduce discriminatory behavior. Striving for egalitarianism—which people generally report as desirable (O'Brien et al., 2010)—requires that people recognize and correct for non-egalitarian behavior. Thus, teaching people to recognize their implicit bias may allow them to reduce their discriminatory behaviors. Some evidence supports this idea; the IAT has been shown to be an effective tool for educating college students about implicit bias (Hillard, Ryan, & Gervais, 2013; Morris & Ashburn-Nardo, 2009). Further, police officers who are trained to recognize their own automatic racial bias are better able to resist shooting an unarmed Black suspect in a shooter game (Correll, Park, Judd, & Wittenbrink, 2002).

However, the present results also suggest a caveat for interventions based on IAT feedback: that the people who may benefit most from awareness of their own implicit attitudes (e.g., majority group members; those whose implicit bias aligns more with societal prejudice than does their explicit bias) also respond most defensively to feedback. As such, interventions based on IAT feedback might avert this caveat via defensiveness-reduction strategies. For instance, they might be more effective by asking participants to think about important values (Steele, 1988), focus on social support (Crocker, Niiya, & Mischkowski, 2008), or engage in dissonance-ameliorating behaviors (Harmon-Jones, 2000).

Directions for Future Studies

It is not clear whether defensiveness is related to people's prior experience with implicit measures. It is possible that people with previous experience with implicit measures would be more, or less, defensive than those with no experience. On one hand, those who already know that their implicit attitudes are biased, or who have learned that most people's implicit attitudes are biased, might have braced for such feedback, leading to *decreased* defensiveness. On the other hand, those who already know that their implicit attitudes are biased, or have learned that most people's implicit attitudes are biased, may begin an IAT already prepared to dismiss the results, leading to *increased* defensiveness. Future studies should examine whether (and how) prior experience with or knowledge about the IAT influences defensive responses.

In Studies 1 and 2, one intriguing finding was that magnitude, direction, group status, and their interactions typically explained more of the variance in defensiveness for evaluative IATs than for stereotyping IATs. This could be because participants perceive evaluative IATs as more threatening to their egalitarian self-concepts. For example, preferring straight people to gay people, or able-bodied people to those with disabilities, clearly indicates that one is not egalitarian toward both groups. By contrast, seeing White people as more American than Asian people does not necessarily indicate that one prefers White individuals to Asian individuals, only that one more readily associates them with the concept American. Moreover, being told that one has a "preference"—the language used in evaluative IAT feedback—may be more threatening than simply being told that one has an "association"—the language used in stereotyping IAT feedback. As such, it would be useful for future inquiries to further examine the role of feedback content in defensiveness across different types of IATs.

In a similar vein, future research is needed to understand whether the effects we observed here extend beyond the procedures we used. Our approach was limited in three

notable ways. First, we only examined responses to the IAT. We chose to do so for three reasons. First, large archival IAT datasets are available, allowing an unusually large, diverse, high-powered test of the hypotheses. Second, reactions to IAT feedback might be especially important because the IAT is the most available measure of implicit attitudes, and is most common for the public to access. Importantly, this allows maximum external generalizability in terms of how most people are likely to hear about, take, and receive feedback from an implicit measure. Still, it is possible that the effects we observed were due to some unique attribute of the IAT process. We assume here that people are responding to their feedback, rather than to taking the IAT. This hypothesis is supported by our finding in Study 3 that IAT feedback, rather than IAT scores influenced participants' defensive responses. Nevertheless, future work can examine the generality of the present findings by replicating the current work with other measures of implicit bias (e.g., the Affect Misattribution Procedure; Payne, Cheng, Govorun, & Stewart, 2005).

Second, we examined explicit beliefs using only a single item that asked people directly about their endorsement of prejudice. A variety of research suggests that when people report their own biases, they do so in a socially desirable way (e.g., McConahay, 1986; Nosek, 2007). Indeed, socially-desirable reporting is one of the reasons that researchers turn to the IAT as an additional measure of attitudes. Moreover, researchers without access to implicit measures typically employ less-direct, multi-item measures of explicit prejudice (e.g., McConahay, 1986). Further, although we infer that people's explicit reports are influenced by egalitarian motives, we did not measure participants' egalitarian motivations directly (e.g., the motivation to respond without prejudice; Plant & Devine, 1998). Given the archival context of the original study, and a desire to keep measures consistent across studies, we measured explicit attitudes using items that directly mirrored the IAT feedback people received. We see this as a strength because explicit attitudes and implicit feedback were therefore easily compared, allowing us to compute magnitude and direction of implicit-explicit discrepancy. Still, future research should attempt to examine whether the present effects replicate using other measures of explicit bias and the specific role of egalitarian motives.

Finally, although our first two studies suggested the importance of discrepancy in defensive responding, our third study did not examine these effects. Due to an unfortunate oversight, we did not include a measure of explicit bias in our third study. As such, we were unable to test for the discrepancy effects we observed in Studies 1 and 2. Nevertheless, data from 1.1 million people who completed the Black-White IAT between 2006 and 2012 at the Project Implicit website suggests that for 96% of participants, learning that they had strong pro-White preference would indicate that their implicit views were more pro-White than their explicit views, with 84% of participants indicating explicit attitudes that were two or more points lower than their feedback (Howell et al., 2015). By contrast, learning they were egalitarian would confirm their explicit views for 55% of participants, with only 17% of participants indicating explicit attitudes that are two or more points discrepant from their feedback. As such, we suspect that our manipulation effectively captured a high- versus low-discrepancy situation. Still, future research is needed to better understand how the effects we observed in Study 3 would change if discrepancy were accounted for.

Contribution and Conclusion

Overall, the present studies represent a significant step in understanding defensive responses to feedback about implicit attitudes. Evidence from three studies suggests that people respond defensively to learning that

their implicit attitudes align with societal bias, especially when their explicit attitudes do not. Moreover, to the extent that people respond defensively to news that they are implicitly biased, they are less willing to engage in behaviors that might change this bias. However, feeling bad about one's feedback seems to provide some compunction, motivating behavioral change.

These results are novel and advance theory in at least five ways. First, our investigation is the first to examine and demonstrate the generalizability of defensive responses across multiple IATs, including both stereotyping and evaluative IATs. Our results suggest that defensive processes may be at play in all IATs, but that factors like group status might moderate these effects. Second, the results contribute to a new, but growing, body of literature focused on actual responses to IAT feedback (Studies 1 and 2), while still demonstrating causal processes (Study 3). Third, this is the first experimental demonstration that several different, and sometimes opposing, responses can occur in response to IAT feedback. Indeed, although other work has considered negative affect and derogation as both being signs of defensiveness, our research suggests that these two reactions can work in opposing ways to inform intentions to engage in egalitarian behavior. Finally, it is the first to show that defensiveness stems both from feedback indicating societally-consistent bias that one does not endorse explicitly (i.e., a magnitude by direction interaction) and also feedback that simply diverges from one's own explicit self view (magnitude, in both directions). Thus, defensiveness does not only stem from learning that implicit attitudes are more "prejudiced" (i.e., one's implicit attitudes are societally-consistent) than one indicated explicitly, but also from learning that one's implicit attitudes are less "prejudiced" (i.e., more societally-inconsistent) than one's explicit attitudes.

In addition to addressing important gaps in the literature, and theoretically expanding knowledge about the IAT, defensiveness, and feedback reception broadly, the present work highlights the need to examine and address defensiveness when attempting to use the IAT as an educational tool. Moreover, the present work offers a promising framework for future studies in the area of defensiveness and IAT feedback. In sum, the present work offers a novel demonstration of an important phenomenon and contributes to knowledge about barriers to bias reduction.

.

References

Abelson, R. P. (1968). Theories of cognitive consistency: a sourcebook.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt implications for motivation and self-regulation in the context of prejudice. *Psychological Science, 18*(6), 524-530.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or" would Jesse Jackson'fail'the implicit association test?". *Psychological Inquiry, 15*(4), 257-278.

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The Rules of Implicit Evaluation by Race, Religion, and Age. *Psychological science*, 0956797614543801.

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

Baumeister, R. F., Dale, K., & Sommer, K. L. (1998). Freudian defense mechanisms and empirical findings in modern social psychology: Reaction formation, projection, displacement, undoing, isolation, sublimation, and denial. *Journal of Personality, 66*(6), 1081-1124. doi: 10.1111/1467-6494.00043

Beasley, C. (1999). *What is feminism?: An introduction to feminist theory*: Sage.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42*(2), 192-212.

Briñol, P., Petty, R. E., & Wheeler, S. C. (2006). Discrepancies between explicit and implicit self-concepts: Consequences for information processing. *Journal of Personality and Social Psychology, 91*(1), 154.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciencies*: Routledge.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*(6), 1314-1329. doi: 10.1037/0022-3514.83.6.1314

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. doi: 10.1037/0022-3514.82.3.359

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.

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive Racism. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology, Vol. 36.* (pp. 1-52). San Diego, CA US: Elsevier Academic Press.

Dunning, D. (2007). Self-image motives: Further thoughts and reflections. *Journal of Consumer Psychology, 17*(4), 258-260. doi: 10.1016/s1057-7408(07)70036-0

Festinger, L. (1962). *A theory of cognitive dissonance* (Vol. 2): Stanford university press.

Frantz, C. M., Cuddy, A. J., Burnett, M., Ray, H., & Hart, A. (2004). A threat in the computer: The race implicit association test as a stereotype threat experience. *Personality and Social Psychology Bulletin, 30*(12), 1611-1624.

Freeman, M. F., & Tukey, J. W. (1950). Transformations Related to the Angular and the Square Root. *Annals of Mathematical Statistics, 21*(2), 305-305.

Gawronski, B., & Bodenhausen, G. V. (2011). The Associative-Propositional Evaluation Model: Theory, Evidence, and Open Questions. *Advances in Experimental Social Psychology, 44*, 59.

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

Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential validity of the implicit association test: comment on Blanton and Jaccard (2006).

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. doi: 10.1037/a0015575 10.1037/a0015575.supp (Supplemental)

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*(3), 1369.

Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology, 67*(3), 451-470.

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*, 1-22.

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

Howell, J. L., Gaither, S. E., & Ratliff, K. A. (2014). Caught in the Middle Defensive Responses to IAT Feedback Among Whites, Blacks, and Biracial Black/Whites. *Social Psychological and Personality Science, Online Ahead of Print*, 1948550614561127.

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., & Shepperd, J. A. (2012). Reducing information avoidance through affirmation. *Psychological Science, 23*(2), 141-145.

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality and Social Psychology, 85*(5), 969.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach*: Routledge.

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale.

McQueen, A., Vernon, S. W., & Swank, P. R. (2013). Construct definition and scale development for defensive information processing: an application to colorectal cancer screening. *Health Psychology, 32*(2), 190-202. doi: 10.1037/a0027311

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

Morris, K. A., & Ashburn-Nardo, L. (2009). The Implicit Association Test as a class assignment: Student affective and attitudinal reactions. *Teaching of Psychology, 37*(1), 63-68.

Nosek, B. A. (2007). Implicit - explicit relations. *Current Directions in Psychological Science, 16*(2), 65-69. doi: Doi 10.1111/J.1467-8721.2007.00477.X

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., . . . Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes.

European Review of Social Psychology, 18, 36-88. doi: Doi 10.1080/10463280701489053

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

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of personality and social psychology, 89*(3), 277.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*(3), 811-832. doi: 10.1037/0022-3514.75.3.811

Pyszczynski, T., & Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression. *Psychological Bulletin, 102*(1), 122-138. doi: 10.1037/0033-2909.102.1.122

Redford, L., & Ratliff, K. A. (2015). Perceived moral responsibility for attitude□based discrimination. *British Journal of Social Psychology*.

Rudman, L. A., Dohn, M. C., & Fairchild, K. (2007). Implicit self-esteem compensation: automatic threat defense. *Journal of Personality and Social Psychology, 93*(5), 798.

Ruiter, R. A., Verplanken, B., Kok, G., & Werrij, M. Q. (2003). The role of coping appraisal in reactions to fear appeals: Do we need threat information? *Journal of Health Psychology, 8*(4), 465-474.

Rydell, R. J., McConnell, A. R., & Mackie, D. M. (2008). Consequences of discrepant explicit and implicit attitudes: Cognitive dissonance and increased information processing. *Journal of Experimental Social Psychology, 44*(6), 1526-1532.

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science, 3*(2), 102-116. doi: 10.1111/j.1745-6916.2008.00068.x

Shepperd, J. A., & Howell, J. L. (2015). Responding to Psychological Threats with Deliberate Ignorance: Causes and Remedies. In P. J. Carroll, R. M. Arkin & A. Wichman (Eds.), *Handbook of Personal Security*. New York, NY: Taylor & Francis.

Sherman, D. K. (2013). Self-affirmation: Understanding the effects. *Social and Personality Psychology Compass, 7*(11), 834-845.

Sherman, D. K., Nelson, L. D., & Steele, C. M. (2000). Do messages about health risks threaten the self? Increasing the acceptance of threatening health messages via self-affirmation. *Personality and Social Psychology Bulletin, 26*(9), 1046-1058.

Shoda, T. M., McConnell, A. R., & Rydell, R. J. (2014). Having explicit-implicit evaluation discrepancies triggers race-based motivated reasoning. *Social Cognition, 32*(2), 190-202.

Swann, W. B., Jr. (1990). To be adored or to be known? The interplay of self-enhancement and self-verification. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior, Vol. 2.* (pp. 408-448). New York, NY US: Guilford Press.

Swann, W. B., Jr., Rentfrow, P. J., & Guinn, J. S. (2003). Self-verification: The search for coherence. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity.* (pp. 367-383). New York, NY US: Guilford Press.

Sweeny, K., Melnyk, D., Miller, W. A., & Shepperd, J. A. (2010). Information avoidance: Who, what, when, and why. *Review of General Psychology, 14*(4), 340-353. doi: 10.1037/a0021288

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193-210. doi: 10.1037/0033-2909.103.2.193

Voils, C. I., Ashburn-Nardo, L., & Monteith, M. J. (2002). Evidence of prejudice-related conflict and associated affect beyond the college setting. *Group Processes & Intergroup Relations, 5*(1), 19-33.

Watson, D., & Clark, L. A. (1999). The PANAS-X: Manual for the positive and negative affect schedule-expanded form. *Retrieved from:* http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1011&context=psychology_pubs.

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. *Communications Monographs, 59*(4), 329-349.

Witte, K., & Morrison, K. (2000). Examining the influence of trait anxiety/repression,Äêsensitization on individuals,Äô reactions to fear appeals. *Western Journal of Communication (includes Communication Reports), 64*(1), 1-27.

Figure 1. Study 2: Indirect effects models.