Empirical Research Paper

Personality and Social
Psychology Bulletin
1–16

DOI: 10.1177/01461672241229006
journals.sagepub.com/home/pspb
Sage

Following Prejudiced Behavior, Confrontation Restores Local Anti-Bias Social Norms

Anna Haoyang Li[1], Elisabeth S. Noland[1], and Margo J. Monteith[1]

Abstract

Does confronting, or calling out prejudiced statements or behaviors, signal anti-bias norms? The current studies (*N* = 1,308) examined this question by assessing observers' perceptions of descriptive and injunctive anti-bias local norms after a prejudiced comment was confronted. Studies 1 and 2 revealed a restorative function of confrontation: Confrontation of bias expressed toward Mexican people strengthened non-Mexican participants' perceptions of descriptive anti-bias local norms compared to leaving bias unconfronted and restored the perception of injunctive anti-bias local norms to baseline level (i.e., when no bias had occurred). Study 3 demonstrated that the norm-signaling function of confrontation is applicable to anti-Black bias among both Black and White participants. Moreover, observing confrontation of anti-Black bias boosted participants' sense that their identity would be safe in the environment, mediated by their perceptions of anti-bias descriptive and injunctive norms. Together, these findings indicate that confrontation effectively transforms norms in the face of bias.

Keywords

bias confrontation, social norms, prejudice, identity-safety

Received August 3, 2023; revision accepted January 11, 2024

Imagine Alex is attending a meeting on her first day of work, and a coworker refers to a recently recruited Black manager as a "diversity hire." Just as Alex starts wondering whether this workplace is racially biased, another coworker calls out the racially biased comment. What will Alex make of this exchange? When bias occurs, as it so often does in workplace and social contexts (Sue & Spanierman, 2020), will confronting it signal anti-bias norms to those who observe the confrontation? Much research concerning bias confrontation focuses on outcomes at the interpersonal level, such as whether confrontation reduces bias for the person who has been confronted, and how confrontation affects evaluations of the confronter (Chaney & Sanchez, 2018; Hildebrand et al., 2023; Parker et al., 2018; Rattan et al., 2023; Wilton et al., 2018). However, the current research takes a more expansive perspective to investigate the possible norm-signaling function of bias confrontation. Understanding whether bias confrontations signal anti-bias norms is important because perceived norms influence behavior (Cialdini et al., 1990; Crandall et al., 2002; Reno et al., 1993).

We tested whether observing confrontation shapes people's perceptions of anti-bias descriptive and injunctive norms, and whether a lone confronter is sufficient or other people in the situation must affirm a confrontation for it to influence norm perceptions. We also considered the possible role of observers' group membership. If Alex in the aforementioned example were Black (i.e., a target group member), would the situation be experienced differently than if she were White (i.e., a non-target group member)? The occurrence of bias threatens target group members' sense that their identity is safe in the environment (Major & O'Brien, 2005). Past research revealed that bias confrontation can boost identity-safety among minoritized individuals if it is affirmed by others (Hildebrand et al., 2020). We extended this research by investigating whether, following a biased incident, confrontation signals anti-bias local norms to target group members, and whether stronger norm perceptions are associated with increased identity-safety.

People Are Sensitive to Bias-Related Norms

Norms are unwritten but understood rules reflecting what behaviors are common (i.e., descriptive norms) and

[1]Purdue University, West Lafayette, IN, USA

Corresponding Author:
Anna Haoyang Li, Department of Psychological Sciences, Purdue University, 703 3rd Street, West Lafayette, IN 47907, USA.
Email: li4268@purdue.edu

Correction (March 2024): Article updated to correct the figure labels for Figures 1 and 2.

appropriate (i.e., injunctive norms) in an environment (Cialdini et al., 1990; Reno et al., 1993). Norm perception has a powerful influence on social behavior (Miller & Prentice, 2016; Paluck et al., 2016; Tankard & Paluck, 2016). In the context of bias and prejudice anti-bias descriptive norms indicate that bias is uncommon, and injunctive anti-bias injunctive norms indicate that bias is unacceptable.

People adjust their attitudes and behavior according to salient norms concerning bias and prejudice. For instance, exposure to peers' opinions about a target group, which conveys descriptive norms, causes people to align their opinions with those of their peers (Blanchard et al., 1994; Crandall et al., 2002; Monteith et al., 1996; Stangor et al., 2001). Environmental cues can signal norms, as in the research of Murrar et al. (2020) using posters and videos to convey inclusiveness. Anti-prejudice norms may also be detected from non-verbal cues, such as when people react to a sexist statement with silence (Koudenburg et al., 2021). In other research, when a male ally expressed support for gender equality, women perceived that the company endorsed gender-equality norms, and they anticipated greater workplace support and respect and less isolation and hostility (Moser & Branscombe, 2022).

In sum, research indicates that social norms about bias and prejudice can be cued and have downstream consequences for curbing bias and encouraging feelings of safety and belonging among marginalized groups.

Will Confrontation Convey Anti-Bias Norms?

Whether bias confrontation will signal egalitarian norms and have downstream consequences is not straightforward. With confrontation, although someone speaks out against bias, they do so after someone else already felt free to express bias. What is an observer to make of such a situation?

A key tenet in social psychology is that people seek causes once something has happened (Heider, 1958). As an uncommon response to bias (Ashburn-Nardo et al., 2008; Shelton & Stewart, 2004; Swim & Hyers, 1999), third-party observers may explain bias confrontation in terms of internal attributes of the confronter (Kelley, 1967) (e.g., "This person is overly sensitive"). Indeed, prior research shows that confronting others comes with social costs, or negative interpersonal impressions and evaluations of confronters levied by people who have been confronted (e.g., Alt et al., 2019; Czopp et al., 2006). Although factors moderate the magnitude of social costs (e.g., greater when targets of bias than allies confront, and when in-group members confront; Drury & Kaiser, 2014; Kutlaca et al., 2020; Schultz & Maddox, 2013), confronters are consistently evaluated negatively and as complainers compared to when bias is not confronted. If confrontation is understood with reference to the confronter's dispositional character only, observers are unlikely to perceive anti-bias environmental norms following a confrontation.

However, observers also look to situational factors to explain behavior (Kelley, 1967), and inferring that anti-bias social norms characterize an environment provides a plausible situational explanation for confrontation (e.g., "This place does not tolerate bias"). Why might people infer anti-bias norms from confrontation? At least in the United States, there are strong societal norms against racial bias (Crandall et al., 2002; Glick & Fiske, 1996). When confrontation occurs in a local context, and it concerns bias that is normatively inappropriate in broader society, people may conclude that the broader societal norms apply in the local environment.

In fact, confrontation may be particularly effective at conveying injunctive norms that bias is not acceptable in the local environment. In a series of studies that examined littering behavior in natural settings, Cialdini et al. (1990, Study 4) used swept (versus unswept) litter to signal disapproval of littering. Seemingly trivial as the manipulation was, participants registered the swept litter as an injunctive norm cue against littering and littered less than the unswept litter condition. Likewise, confrontation may serve to "clean up" an environment that has been tainted by a biased remark by activating an injunctive social norm against bias. In addition, whether people infer a local anti-bias descriptive norm after observing a bias confrontation may depend on how other people in the situation respond. If other people speak out to affirm a confrontation (i.e., high consensus; Kelley, 1967), people may be more likely to infer that most people in this setting do not engage in biased behavior than seeing only one person speak out.

The Current Research

Three studies investigated whether, when a biased statement is confronted in a local context, observers conclude that bias is neither common (descriptive norms) nor condoned (injunctive norms) in that environment, relative to a no-bias condition (all studies) and a condition in which bias occurred but was not confronted (Studies 2 and 3). Studies 1 and 2 examined non-Hispanic/Latinx participants' perceptions of anti-bias norms in a company setting after they listened to an audio recording where an anti-Mexican biased remark was confronted by a White male, and other people in the company either affirmed or did not affirm the confrontation. Study 3 replicated Study 2 but concerned anti-Black bias and included both Black and White participants.

Studies 2 and 3 also enriched our understanding of the effects of observing a confrontation by assessing other outcomes that could be expected to co-vary with norm perceptions (i.e., expectations that bias would be sanctioned at the company, future intentions to monitor biases, social costs directed at the confronter, and attributions for the confrontation). Moreover, Study 3 extended previous research showing that affirmed confrontations can boost identity-safety in the face of bias among minoritized group members

Table 1. Participant Demographic Information, Studies 1–3.

	Study 1	Study 2	Study 3	
Demographic Characteristics	N = 342	N = 404	N = 562	
Racial/ethnic group identification				
African American/Black	10.2%	8.2%	52.1%	
Asian/Asian American	8.5%	6.9%		
Caucasian/White	76.3%	80.2%	47.9%	
Biracial	1.8%	2.0%		
Multiracial	1.8%	1.7%		
Remaining options	<1%	<1%		
Gender identification			White Ps	Black Ps
Man	36.5%	37.6%	36.8%	36.9%
Woman	62.0%	59.7%	61.7%	62.1%
Remaining options	1.2%	2.6%	1.5%	1.0%
Age: *M (SD)*	44.50 (14.43)	43.76 (13.97)	43.35 (12.64)	38.60 (13.24)
Political identification: *M (SD)* (1 = *very liberal*; 4 = *neutral*; 7 = *very conservative*)	3.49 (1.80)	3.68 (1.78)	3.65 (1.91)	3.47 (1.54)

Note. Remaining options for racial/ethnic identity were Middle Eastern (Arab or non-Arab), Native American, and "A different identity." Remaining options for gender were transgender woman, transgender man, non-binary, genderfluid, and "I prefer a different term."

(Hildebrand et al., 2020). Specifically, participants' sense of identity-safety was also assessed following the confrontation manipulation, and we tested whether anti-bias norm perceptions statistically mediated the effect of confrontation on identity-safety.

Participants in earlier studies were excluded from participating in later studies. We report how sample sizes were determined and all data exclusions, manipulations, and measures. Materials, data, and Supplemental Material (SM) are available at https://osf.io/a8926/?view_only=47ec5bf9299947b3988e670157491161. All studies were pre-registered: Study 1, http://bit.ly/3Qfdf0B; Study 2, https://bit.ly/3DCSCnw; Study 3, https://bit.ly/45awdK8.

Study 1

We examined the effects of observing an affirmed versus non-affirmed bias confrontation in a particular environment on peoples' perceptions of injunctive and descriptive anti-bias norms in that environment and compared these effects with a control condition where no bias was expressed or confronted. We expected confrontation to communicate anti-bias injunctive norms, predicting that perception of injunctive norms would be stronger in confrontation conditions than in the no-bias condition. We did not have a prediction for the comparison of the affirmed and non-affirmed confrontation conditions. As a noticeable critique of biased behavior, any confrontation may activate injunctive norms to the same extent, although perhaps affirmation would provide a boost. In contrast, we predicted that affirmation would boost the perception of descriptive norms, relative to a non-affirmed confrontation and the no-bias control condition. With others speaking up to join the confronter, we expected greater perception that most people avoid biased behavior.

Method

Participants and Design. We used a single-factor (confrontation condition: no bias, non-affirmed confrontation, affirmed confrontation) between-participants design. An a priori G*Power (Faul et al., 2009) analysis for a one-way ANOVA using a small-medium effect size estimate (f = 0.18), 80% power, and α = .05 suggested 303 participants.

Participants were 342 non-Hispanic/Latinx U.S. Amazon Mechanical Turk workers (MTurk; paid $1). Table 1 provides demographic information for all studies. We excluded one additional participant who did not give post-session consent and retained two participants who failed attention checks because their exclusion did not change results. A sensitivity analysis indicated that our sample size provided 80% power to detect an effect size of f = 0.17 at α = .05 (equivalent to η_p^2 = .03).

Procedure. After consenting to participate, participants learned that the study examined people's perceptions of new environments and were informed that they would listen to an audio of four people interacting in a work setting. Before the audio started, participants read a short description indicating that the setting was a breakroom at a branch office of a 20-year-old retail e-commerce company, Beier Inc. Participants read that there are about 200 employees at this branch office and that they are at the hiring period of their annual recruiting cycle. Participants were shown pictures, names, and ages of four coworkers interacting in the audio.

Participants were randomly assigned to one of the three confrontation conditions. All participants listened to a 3-minute conversation among the four coworkers that was identical until the last part when the conversation turned to hiring (audio scripts adapted from Hildebrand et al., 2020, Studies 2 and 3).

In the *no-bias* confrontation condition, the discussion around hiring did not include a biased statement or a confrontation.

In the *non-affirmed* confrontation condition, a White man made the following biased statement: "Honestly, I'm just not sure if there'd be a qualified candidate who's Mexican. IT is central to the company's functioning and requires a lot of advanced qualifications and brainpower . . .." Then another White male confronter said,

> Woah, let's just backtrack for a moment. If we do have any Mexican candidates just know that they are as qualified and capable for this job as anyone else! And also, I had said there are two *Hispanic* candidates. Hispanic people are not just Mexican. The term Hispanic refers to any people from a Spanish-speaking country.

The same biased comment and confrontation occurred in the *affirmed* confrontation condition, but immediately after, the two other people involved in the conversation (one White and one Asian woman) chimed in: "I agree. Let's not assume that Mexican people make poor directors," and "Yeah, race isn't an indicator of how qualified a person is." Data from a separate MTurk sample (N = 50) provided evidence that these comments were interpreted as strong affirmations of the confrontation. Specifically, participants' ratings on seven-point (*not at all* to *very much*) scales indicated that they thought the women's comments conveyed that they endorsed the confrontation and agreed with it, averaged to form composite, $r = .75$; $M = 5.93$, $SD = 1.31$; one-sample t-test with 4.0 (scale midpoint) test value: $t(49) = 10.43$, $p < .001$. Also, participants reported that the women who affirmed the confrontation thought the biased comments were troublesome and unacceptable, averaged to form composite, $r = .58$; $M = 5.95$, $SD = 1.35$; one-sample t-test with 4.0 test value: $t(49) = 10.22$, $p < .001$.

In all conditions, the audio ended as the conversation turned to a different topic and immediately faded out. Participants then completed measures in the following order, followed by a post-session consent form that disclosed the full study purpose and asked for permission to use their data.

Measures

Perceptions of Descriptive Norms. Consistent with the definition of descriptive norms and how they are typically conveyed (Blanton et al., 2008; Reno et al., 1993), we assessed descriptive norm perceptions by asking participants what percentage of people at Beier Inc. engaged in certain behaviors on 0% to 100% sliding scales. Three of 10 items assessed perceptions of anti-bias descriptive norms: "What percentage of employees at this workplace never say negative things about Mexican people while at work?," "What percentage of employees at this workplace feel free to question the intelligence of Mexican people while at work?" (reverse-scored), and "What percentage of employees at this workplace are comfortable using negative stereotypes of Mexican people while at work?" (reverse-scored). Composite scores of perceptions of descriptive norms were calculated by averaging participants' responses to the three critical items ($M = 73.69$, $SD = 19.26$; $\alpha = .65$[1]).

Perceptions of Injunctive Norms. Injunctive norms were measured with items assessing approval/disapproval of relevant behavior (e.g., Baer, 1994). Accordingly, participants completed 10 items, rating each on a 1 (*strongly disagree*) to 9 (*strongly agree*) scale, to indicate the extent to which certain behaviors in this workplace were approved or disapproved. Three critical items assessed perceptions of anti-bias injunctive norms: "Saying anything negative about Mexican people at this workplace is strongly disapproved," "At this workplace, it is completely unacceptable to put down the intelligence of Mexican people in any way," and "At this workplace, employees should never use any negative stereotypes about Mexican people." Due to a programming error in Study 1 only, the third item could not be used, and composite scores were calculated by averaging participants' responses to the first two critical items, $M = 7.08$, $SD = 2.00$; $r(340) = .69$, $p < .001$.[2]

Results

One-way between-participant analyses of variance (ANOVAs) were performed on each measure. Given a priori hypotheses for most cell comparisons, Fisher's least significant difference tests were used for pairwise comparisons. For all studies, we report effect size ds for significant pairwise comparisons, followed by 95% confidence intervals around these effect sizes in brackets.

Perceptions of Descriptive Norms. The main effect for confrontation condition on perceptions of descriptive norms was significant, $F(2, 339) = 7.11$, $p < .001$, $\eta_p^2 = .04$. As shown in Figure 1 (Panel A), when no bias was expressed in the first place, participants assumed most people do not engage in bias at this workplace. Once bias occurred, participants perceived fewer people in this environment never do biased things. Specifically, compared to the no-bias condition, participants perceived weaker anti-bias descriptive norms in both the non-affirmed, $p < .001$, $d = 0.47$ [0.21, 0.73], and affirmed, $p = .008$, $d = 0.36$ [0.09, 0.62], confrontations. The two confrontation conditions did not differ significantly, $p = .382$. What we do not know from these data is whether confrontation may increase anti-bias descriptive norms compared to a situation in which bias occurs but is not confronted.

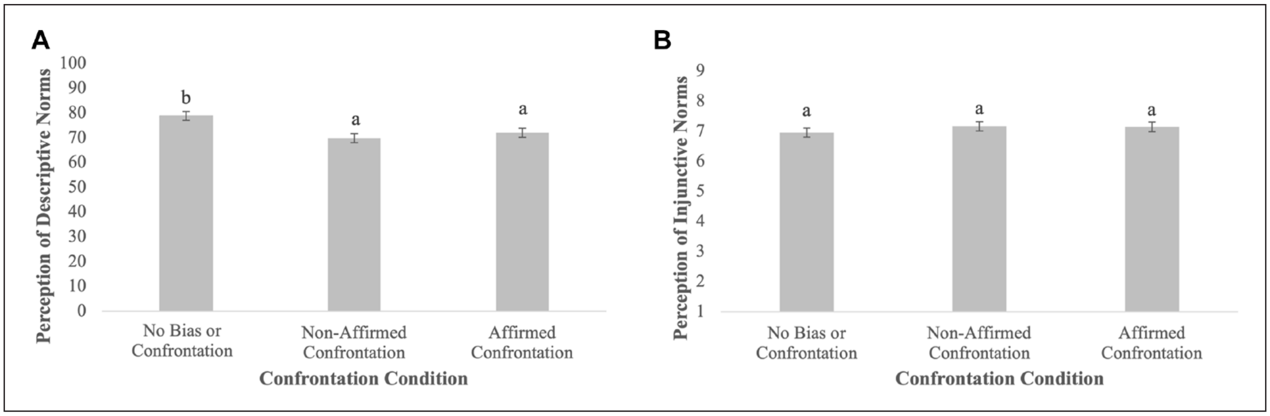


Figure 1. Descriptive (Panel A) and Injunctive (Panel B) Norm Perceptions as a Function of Confrontation Condition, Study 1.
Note. Error bars represent ±1 standard error. Means not sharing a lowercase letter differ significantly from each other.

The finding that anti-bias descriptive norm ratings were well above the midpoint in the confrontation conditions is consistent with this possibility.

Perceptions of Injunctive Norms. Contrary to hypotheses, the effect of confrontation condition on perceptions of injunctive norms was not significant, F (2, 339) = 0.56, p = .573, η_p^2 = .003 (Figure 1 [Panel B]). Remarkably, participants perceived equally strong anti-bias injunctive norms whether bias did not occur in the first place or bias occurred and then was confronted and either affirmed or not. One interpretation of these findings is that confrontation restores an anti-bias injunctive norm to a level when no bias occurs in the first place. However, a condition in which bias occurs but is not confronted is needed to test whether this is the case.

Discussion

Study 1 results were intriguing: Although they did not conform to our hypotheses that bias confrontation would generate stronger anti-bias norm perceptions relative to a condition free of bias, the patterns of results may suggest a restorative function of confrontation. That is, if people generally presume strong norms against bias when no bias occurs, can confrontation undo the damage caused by the bias incident and re-establish the default perceptions that bias is neither common nor condoned? We tested this question in Study 2.

Study 2

Study 2 tested the same hypotheses as in Study 1. In addition, we added a condition in which bias was expressed but not confronted to test how much confrontation strengthens the perception of anti-bias norms in the face of bias. We predicted that anti-bias descriptive and injunctive norm perceptions would be weakest when bias was not confronted and significantly stronger when bias was confronted.

Study 2 also assessed whether participants expected sanctions for bias in the local environment, and whether they would monitor their own biases in the environment. We predicted that expected sanctions and monitoring intentions would be stronger when bias was confronted than not, consistent with the ideas that injunctive norm-violating behavior comes with sanctions (Cialdini & Trost, 1998) and that confrontations can encourage people to self-regulate their own biases (Monteith et al., 2022).

Other measures concerned how observers perceived the confronter and the confrontation. Specifically, social costs directed at the confronter were assessed (e.g., complainer). We predicted that greater social costs would be directed at a person when they confronted than when they did not, consistent with prior research (e.g., Chaney & Sanchez, 2018; Hildebrand et al., 2023). We also predicted greater social costs in the non-affirmed than affirmed confrontation condition, reasoning that the support for the confrontation provided through consensus would discourage negative judgments about the confronter. Finally, we assessed dispositional and situational attributions for the confronter's behavior in the confrontation conditions. Because affirmation conveys consensus information, which encourages situational attributions (Kelley, 1967), we predicted weaker dispositional than situational attributions in the affirmed confrontation condition. In contrast, we anticipated stronger dispositional than situational attributions in the non-affirmed confrontation condition.

Method

Participants and Design

We used a single-factor (confrontation condition: no bias, bias not confronted, non-affirmed confrontation, affirmed confrontation) between-participants design. An a priori G*Power analysis for a one-way ANOVA with an effect size of f = 0.18 indicated 344 participants would provide 80% power at α = .05. We recruited 404 non-Hispanic/Latinx U.S. participants from MTurk (paid $1). Two additional participants were excluded because they did not give

Table 2. Inter-Measure Correlations, Study 2.

Measure	1	2	3	4	5	6
Descriptive norm perceptions	–	–	–	–	–	–
Injunctive norm perceptions	.47**	–	–	–	–	–
Expected sanctions for bias	.48**	.66**	–	–	–	–
Future intentions to monitor biases	.03	.10	.14**	–	–	–
Social costs	−.20**	−.22**	−.28**	−.16**	–	–
Dispositional attribution[a]	.17*	.09	.28**	.29**	−.45**	–
Situational attribution[a]	−.19**	−.01	.00	.02	.07	.14*

[a]Attribution items were completed in the confrontation conditions only.
*$p < .05$. **$p < .01$.

post-session consent. A sensitivity analysis indicated that our final sample size had 80% power to detect an effect size of $f = .17$ at $\alpha = .05$ (equivalent to $\eta_p^2 = .03$).

Procedure. The procedure replicated Study 1 but added a condition where a biased remark was not confronted and included additional measures. In the bias-not-confronted condition, the biased remark was followed by a 2-second pause and then the conversation shifted to a different topic and faded out. Participants completed measures in the same order as they are presented in the following sections, except that the measures of descriptive and injunctive norms were counterbalanced.

Measures

Descriptive and Injunctive Norms. The measures of perceptions of descriptive ($M = 73.31$, $SD = 19.19$, $\alpha = .64$) and injunctive ($M = 6.80$, $SD = 2.23$, $\alpha = .92$) norms were identical to those of Study 1.

Future Intentions to Monitor Biases. Participants imagined they started a position at the company and completed eight items concerning whether they would monitor certain behaviors at work, using a 1 (*very unlikely*) to 9 (*very likely*) scale. Composite scores of future intentions to monitor biases against Mexican people were calculated by averaging participants' responses to three critical items (e.g., "I would be on guard so that stereotypes about Mexican people never affect what I say or do," $M = 7.00$, $SD = 1.83$, $\alpha = .65$).

Attributions. Only participants in the confrontation conditions completed attribution items. They were reminded of what the confronter, Dan, said in the audio and rated two items on a 1 (*strongly disagree*) to 7 (*strongly agree*) scale: "The kind of person Dan is (such as his character, his attitudes, or temperament) influenced his behavior" (dispositional attribution) and "The kind of situation Dan was in (such as the atmosphere, social norms, or other contextual factors) influenced his behavior" (situational attribution).[3]

Social Costs. Participants rated Dan, who confronted bias in the confrontation conditions only, on 12 traits (1 = *does not apply at all*, 7 = *applies very much*), including six critical traits that were averaged (Czopp et al., 2006; $\alpha = .89$; e.g., "hypersensitive," "hostile") ($M = 2.52$, $SD = 1.22$). Participants also evaluated Dan on five items (1 = *strongly disagree*, 7 = *strongly agree*) (Mallett & Wagner, 2011; $\alpha = .95$; e.g., "I feel negatively toward Dan," "I would like to hang out with Dan" [reverse-scored]) ($M = 3.07$, $SD = 1.54$). As preregistered and consistent with the study by Hildebrand et al. (2023), the measures were standardized within their respective distributions, and scores were averaged to create a social costs index (reliability of the composite = .95; Nunnally, 1978). Higher scores indicate greater social costs.

Expected Sanctions for Bias. Participants rated the likelihood of eight outcomes happening at the workplace using a 1 (*very unlikely*) to 7 (*very likely*) scale. Ratings for three critical items (e.g., "An employee who tells racist jokes about Mexican people at this company will receive disciplinary action from the Human Resources department") were averaged ($M = 5.19$, $SD = 1.41$, $\alpha = .80$).

Results

One-way between-participants ANOVAs were performed on all dependent variables unless noted otherwise. See Table 2 for measure intercorrelations.

Perceptions of Descriptive Norms. The main effect of confrontation condition was significant, $F(3, 400) = 11.78$, $p < .001$, $\eta_p^2 = .08$, see Figure 2 (Panel A). As expected, participants perceived significantly weaker anti-bias descriptive norms when bias occurred but was not confronted, relative to the no-bias condition, $p < .001$, $d = 0.75$ [0.46, 1.04]. Compared to leaving bias unconfronted, the perception of anti-bias descriptive norms was stronger for both non-affirmed, $p < .001$, $d = 0.46$ [0.18, 0.74], and affirmed, $p = .031$, $d = 0.29$ [0.01, 0.56], confrontations. As in Study 1, the non-affirmed and affirmed confrontation conditions did not differ, $p = .227$. Descriptive norm perceptions were still weaker in the non-affirmed, $p = .012$, $d = 0.38$ [0.10, 0.66], and affirmed, $p < .001$, $d = 0.53$ [0.25, 0.81], confrontation conditions than when bias did not occur in the first place.

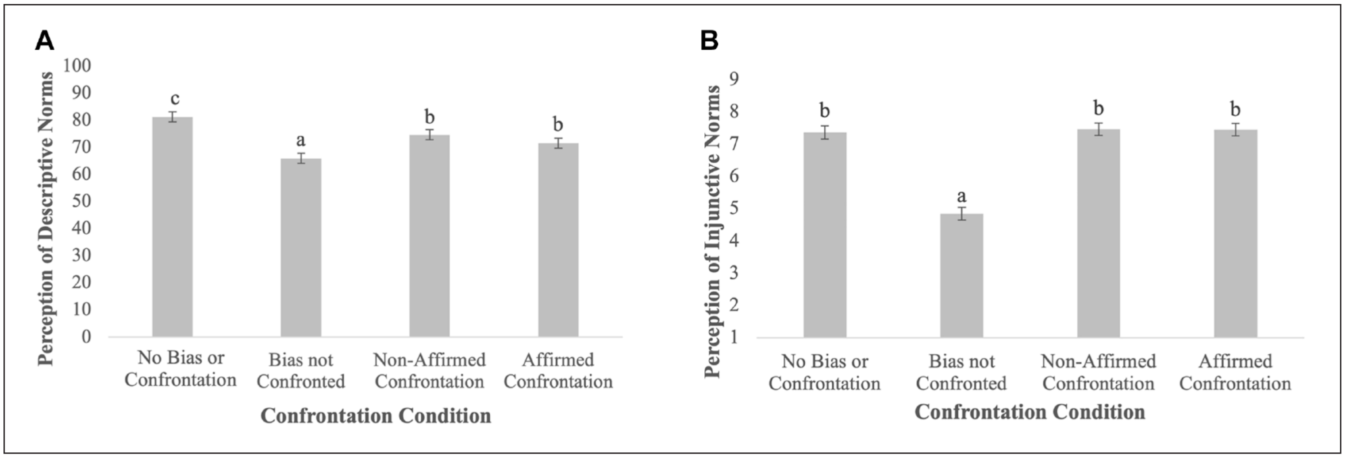


Figure 2 Descriptive (Panel A) and Injunctive (Panel B) Norm Perception as a Function of Confrontation Condition, Study 2
Note. Error bars represent ±1 standard error. Means not sharing a lowercase letter differ significantly from each other.

Table 3. Means (SDs) as a Function of Confrontation Condition, Study 2.

Measure	No bias (n = 99)	Bias not confronted (n = 98)	Non-affirmed confrontation (n = 101)	Affirmed confrontation (n = 106)
Expected sanctions for bias	5.51_{b} (1.29)	4.20_{a} (1.49)	5.51_{b} (1.14)	5.50_{b} (1.24)
Future intentions to monitor biases	6.76_{ab} (1.99)	7.23_{ab} (1.71)	7.26_{b} (1.73)	6.74_{a} (1.84)
Social costs	-0.12_{a} (0.73)	0.25_{b} (0.94)	-0.29_{a} (0.87)	0.15_{b} (1.01)

Note. Within each row, means without a common subscript differ significantly ($p < .05$).

Perceptions of Injunctive Norms. The confrontation condition main effect was significant, $F(3, 400) = 44.02, p < .001, \eta_p^2 = .25$, see Figure 2 (Panel B). Consistent with our hypothesis, the perception of anti-bias injunctive norms was much weaker in the bias-not-confronted condition than in the no-bias condition, $p < .001$, $d = 1.16$ [0.86, 1.46]. However, when bias was confronted, with or without affirmation, participants perceived anti-bias injunctive norms to be just as strong as in the no-bias condition, ps $> .70$. The non-affirmed and affirmed conditions did not differ, $p = .963$.

Expected Sanctions for Bias. A significant confrontation condition main effect, $F(3, 400) = 25.42, p < .001, \eta_p^2 = .16$, had the same pattern as injunctive norm perceptions (see Table 3). Participants expected significantly weaker workplace sanctions for bias when bias was not confronted than for the no-bias condition, $p < .001$, $d = 0.94$ [0.65, 1.23]. Furthermore, affirmed and non-affirmed bias confrontations produced expected sanctions that were greater than those in the bias-not-confronted condition, $p < .001$, $d = 0.95$ [0.66, 1.24], $p < .001$, $d = 0.99$ [0.69, 1.28], respectively, and that were similar to the level of expected sanctions in the no-bias condition, ps $> .96$

Future Intentions to Monitor Biases. The confrontation condition main effect was trending, $F(3, 400) = 2.54$, $p = .056$, $\eta_p^2 = .02$. Post hoc analyses did not support our predictions. As shown in Table 3, intentions to monitor bias were somewhat ($p = .069$) stronger after observing bias that was not confronted than for the no-bias condition. Monitoring intentions were also somewhat stronger after observing a lone confronter than the no-bias condition, $p = .050$, $d = 0.27$ [0.01, 0.55], and the affirmed confrontation condition, $p = .039$, $d = 0.29$ [0.02, 0.57]. Before making too much of these unexpected patterns, we tested for replication in Study 3.

Social Costs. The confrontation condition main effect was significant, $F(3, 400) = 7.54, p < .001, \eta_p^2 = .05$ (see Table 3). Contrary to predictions, participants levied greater social costs toward Dan when he failed to confront bias, relative to the no-bias condition, $p = .004$, $d = 0.43$ [0.15, 0.71]. In other words, having just witnessed bias, observers disliked a person who did not stand up to the bias. Also contrary to predictions, social costs were greater in the affirmed than the non-affirmed confrontation condition, $p < .001$, $d = 0.34$ [0.07, 0.62]. Together, these results suggest that observers of a bias incident evaluated a person who did not speak out against the bias negatively, whereas they rendered more favorable evaluations toward a person who confronted without the support of others.

Attributions. A mixed model, 2 (affirmed vs. non-affirmed confrontation) × 2 (dispositional vs. situational attribution; within-participants variable) ANOVA revealed a large main effect for attribution type, $F(1, 205) = 105.23$, $p < .001$, $\eta_p^2 = .34$. Unexpectedly, participants overwhelmingly made stronger dispositional ($M = 5.91$, $SD = 1.36$) than situational ($M = 4.29$, $SD = 2.06$) attributions. A much smaller interaction effect also emerged, $F(1, 205) = 4.70$, $p = .03$, $\eta_p^2 = .02$. Participants made stronger dispositional than situational attributions in the affirmed ($M_{dispositional} = 5.59$, $SD = 1.36$; $M_{situational} = 4.31$, $SD = 2.00$) and non-affirmed ($M_{dispositional} = 6.24$, $SD = 1.20$; $M_{situational} = 4.27$, $SD = 2.13$) confrontation conditions; however, the difference was attenuated in the affirmed, $F(1, 205) = 33.54$, $p < .001$, $\eta_p^2 = .14$, compared to the non-affirmed, $F(1, 205) = 75.37$, $p < .001$, $\eta_p^2 = .27$, condition. This attenuation is consistent with the expectation that consensus would weaken dispositional attributions, although we had predicted a full reversal (i.e., stronger situational than dispositional attributions with affirmation).

Discussion

Study 2 sheds light on Study 1 findings based on the inclusion of a condition in which bias was not confronted. Across the studies, we can conclude that bias confrontation signals both anti-bias descriptive and injunctive norms. The null effect for affirmation is interesting: Just one person confronting bias has the power to influence norm perceptions among non-target group members.

Other findings were also consistent with a shift in norm perceptions. First, as predicted, participants expected greater sanctions for expressing bias when bias had been confronted versus not confronted. Indeed, expected sanctions were just as strong with confrontation as when bias had not occurred in the first place. This pattern aligns with our injunctive norm results and the broader literature, suggesting that injunctive norms discourage counter-normative behavior by promising social sanctions (e.g., Cialdini et al., 1990). Second, participants' intentions to monitor their own bias in the environment were greater with the non-affirmed confrontation than with the no-bias condition, although oddly this was not the case in the affirmed confrontation condition.

Study 2 also examined participants' perceptions of the confronter, Dan. Contrary to our hypotheses, observers evaluated Dan more negatively when he did not confront bias than when no bias was expressed and when he alone confronted bias. Although researchers frequently find that confronters are evaluated negatively by people who are confronted (e.g., Czopp et al., 2006), perhaps confronters of others' bias are regarded positively.

Finally, participants made stronger dispositional than situational attributions for Dan's confronting behavior whether the confrontation was affirmed or not, which was unexpected, although the finding that dispositional attributions were attenuated when confrontation was affirmed (i.e., with consensus) is consistent with Kelley's (1967) covariation model. Given the attribution results do not make a critical contribution to this research program, we decided not to assess attributions in Study 3.

Study 3

Study 3 investigated bias confrontation as a norm-signaling cue for anti-Black bias among both Black and White participants. Members of historically disadvantaged groups are vigilant for bias-relevant cues (Cheryan et al., 2009; Major et al., 2003). The threat of being judged through negative stereotypes creates unsafe situations for targets of bias, thereby generating distrust of the setting and disinterest in joining it (e.g., Murphy et al., 2007; Purdie-Vaughns et al., 2008). After a biased comment, will confrontation repair the harm? We expected confrontation to increase Black participants' perceptions of anti-bias norms and to boost identity-safety and that the effect of confrontation on identity-safety would be statistically mediated by the perceptions of anti-bias norms. Given previous findings that identity-safety among targets of bias was boosted only if confrontation was affirmed by others (Hildebrand et al., 2020), we expected these effects to be observed primarily in the affirmed confrontation condition.

Our hypotheses regarding White participants' norm perceptions remained the same as Study 2. Although not of critical interest, we hypothesized that Black participants overall would perceive weaker anti-bias social norms than White participants. Because we were mainly interested in how bias confrontations affect Black participants' identity-safety, we had no a priori hypotheses for White participants on this dependent variable.

As in Study 2, we also assessed expected sanctions for bias, intentions to monitor bias (White participants only), and social costs. We had the same hypotheses for these measures as in Study 2 but additionally expected that, overall, Black participants would expect less severe sanctions than White participants, and the affirmed confrontation would increase expected sanctions more than the non-affirmed confrontation among Black participants.

Method

Participants and Design. We used a 2 (participant race: White vs. Black) × 4 (confrontation condition: no bias, bias not confronted, non-affirmed confrontation, affirmed confrontation) between-participants design. G*Power revealed that 341 participants would provide 80% power at $\alpha = .05$ with a small-medium effect size estimate, $f = .18$. However, given interaction effects are often difficult to detect and require more participants than suggested by G*Power (Giner-Sorolla, 2018), data were collected from 640 U.S. participants from MTurk (paid $1) and Prime Panels (variable cost, $M = \$5.03$).[4] After pre-registered exclusions (34 did not identify with being either Black or White; 44 failed attention checks and results differed slightly with their exclusion), 562 participants

Table 4. Inter-Measure Correlations, Study 3.

Measure	α	1	2	3	4	5	6
Descriptive norm perception	.69	–	–	–	–	–	–
Injunctive norm perception	.90	.34**	–	–	–	–	–
Identity-safety	.91	.40**	.42**	–	–	–	–
Expected sanctions for bias	.66	.45**	.54**	.56**	–	–	–
Future intentions to monitor biases	.54	.02	.09	−.11	.06	–	–
Social costs	–	−.33**	−.15**	−.42**	−.21**	−.01	–

Note. Reliability for expected sanctions and intentions to monitor biases increased when the reverse-scored items in each index are removed (α = .81 and .66, respectively), but analyses of the two-item indexes yielded results very similar to those reported in the text.
$^{*}p < .05$. $^{**}p < .01$.

remained. A sensitivity analysis indicated that our final sample size had 80% power to detect an effect size of $f = .14$ at $\alpha = .05$ (equivalent to $\eta_p^2 = .02$, or a small effect).

Procedure. We replicated Study 2, except (a) participants read rather than listened to the workplace exchange; (b) candidates were being considered for a software developer position; (c) the biased comment made by a White man was, "I don't understand how this Black guy made it on the shortlist . . . I mean, I've never thought of Black people as good at software development. Would this be a diversity hire to make us look good or what?"; (d) in the confrontation conditions, another White man then said, "C'mon, what are you suggesting here? Deshawn's application is pretty strong! Being a great software developer has nothing to do with being Black. You're not being fair"; and (e) the affirmations referred to Black people.

Measures. We used the same measures as in Study 2, altered to refer to Black people. Only White participants rated intentions to monitor biases toward Black people. The identity-safety measure was new to this study.

Identity-Safety. Identity-safety was assessed with 12 items assessing belonging (adapted from Johnson & Pietri, 2023; eight items; e.g., "I would feel respected at this company") and comfort (adapted from Purdie-Vaughns et al., 2008; four items; e.g., "I think that I could trust my colleagues to treat me fairly at this company"). Ratings were made on a 1 (*strongly disagree*) to 7 (*strongly agree*) scale ($M = 4.00$, $SD = 1.23$, $\alpha = .94$).

Results

Each dependent variable was predicted using a 2 (participant race) × 4 (confrontation condition) between-participant ANOVA, with one exception noted in the following section. Intercorrelations among study measures are shown in Table 4.

Perceptions of Descriptive Norms. As expected, Black participants ($M = 55.21$, $SD = 23.72$) perceived weaker anti-bias descriptive norms than White participants ($M = 67.15$, $SD = 21.78$), $F(1, 554) = 48.11$, $p < .001$, $\eta_p^2 = .08$. A significant main effect of confrontation condition, $F(3, 554) = 24.92$, $p < .001$, $\eta_p^2 = .12$, was qualified by an interaction, $F(3, 554) = 5.84$, $p < .001$, $\eta_p^2 = .03$. As shown in Figure 3 (Panel A), the interaction emerged due to confrontation (non-affirmed and affirmed) creating stronger perceptions of anti-bias descriptive norms, relative to the bias-not-confronted condition, to a greater extent among Black participants than among White participants.

Specifically, anti-bias descriptive norms were perceived as significantly weaker when bias was not confronted, relative to the no-bias condition, among both Black, $p < .001$, $d = 0.88$ [0.53, 1.21], and White, $p < .001$, $d = 0.95$ [0.59, 1.30], participants. Anti-bias descriptive norms were perceived to be stronger with confrontation than when bias was not confronted: The difference was relatively large for Black participants in both the non-affirmed, $p < .001$, $d = 0.59$ [0.25, 0.93], and affirmed, $p < .001$, $d = 1.14$ [0.79, 1.48], conditions. The difference was less pronounced for White participants, unexpectedly not significant for the non-affirmed confrontation, $p = .492$, $d = 0.11$ [−0.22, 0.44], but significant for the affirmed confrontation, $p = .037$, $d = 0.36$ [0.02, 0.70]. In addition, as predicted, the affirmed confrontation produced stronger descriptive norm perceptions than the non-affirmed confrontation among Black participants, $p < .001$, $d = 0.61$ [0.28, 0.94], whereas this difference was not significant for White participants, $p = .155$, $d = 0.25$ [−0.09, 0.59]. Finally, note that Black participants reported that anti-bias descriptive norms were just as strong when bias was confronted as when no bias occurred in the first place, with neither the non-affirmed nor affirmed conditions differing from the no-bias condition, *p*s > .05. In contrast, compared to the no-bias condition, White participants perceived weaker anti-bias descriptive norms for both non-affirmed, $p < .001$, $d = 0.84$ [0.48, 1.19], and affirmed, $p = .002$, $d = 0.62$ [0.26, 0.96], confrontations.

Perceptions of Injunctive Norms. A trending race main effect, $F(1, 554) = 3.60$, $p = .058$, $\eta_p^2 = .01$ (Black participants, $M = 6.62$, $SD = 2.53$; White participants, $M = 6.21$, $SD =$

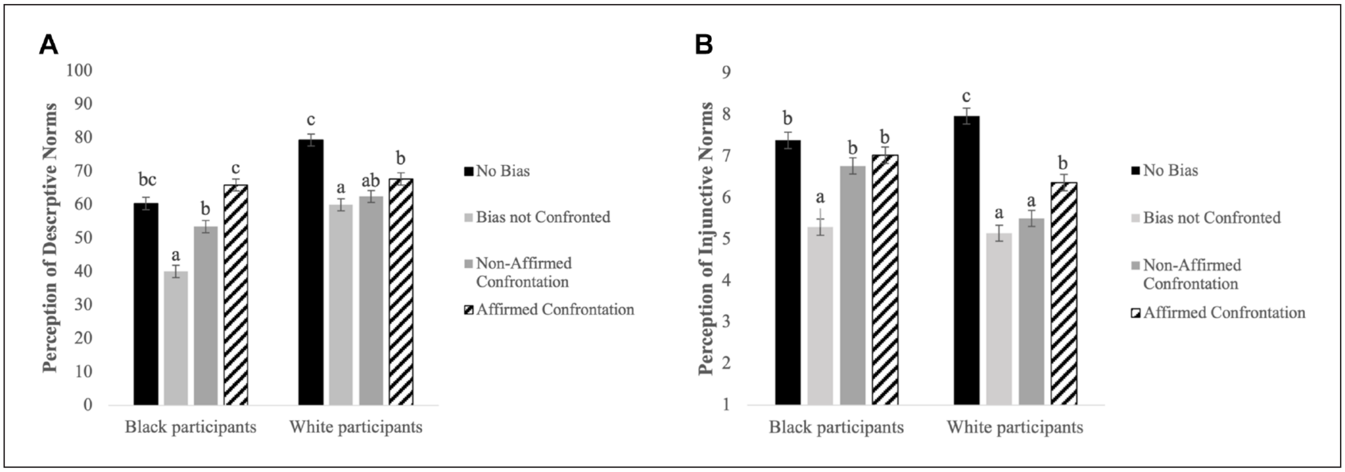


Figure 3. Descriptive (Panel A) and Injunctive (Panel B) Norm Perception as a Function of Confrontation Condition, Study 3.
Note. Error bars represent ±1 standard error. Within race, means not sharing a lowercase letter differ significantly from each other.

2.41), and a significant main effect of confrontation condition, $F(3, 554) = 27.74$, $p < .001$, $\eta_p^2 = .13$, were qualified by a significant interaction between race and confrontation condition, $F(3, 554) = 4.00$, $p = .008$, $\eta_p^2 = .02$ (see Figure 3 [Panel B]). The pattern of the interaction was nearly identical to descriptive norms and driven by the greater impact of confrontation among Black than among White participants.

Specifically, weaker anti-bias injunctive norms were perceived when bias was not confronted than in the no-bias condition among Black, $p < .001$, $d = 0.80$ [0.46, 1.14], and White, $p < .001$, $d = 1.35$ [0.96, 1.71], participants. Among Black participants, both non-affirmed, $p < .001$, $d = 0.53$ [0.19, 0.86], and affirmed, $p < .001$, $d = 0.66$ [0.33, 0.98], confrontation caused participants to perceive stronger anti-bias injunctive norms than when bias was not confronted. Among White participants, the non-affirmed confrontation condition unexpectedly did not differ from the bias-not-confronted condition, $p = .357$, whereas participants in the affirmed confrontation condition perceived stronger anti-bias injunctive norms than those in the bias-not-confronted condition, $p = .002$, $d = 0.51$ [0.17, 0.86]. The non-affirmed and affirmed conditions were comparable among Black participants, $p = .497$, whereas White participants reported stronger injunctive norm perceptions in the affirmed than non-affirmed condition, $p = .031$, $d = 0.38$ [0.04, 0.72]. Finally, Black participants' perceptions of anti-bias injunctive norms were just as strong when bias was confronted, whether affirmed or not, as when no bias occurred in the first place, *p*s > .11. However, compared to the no-bias condition, White participants' perceptions were weaker for both non-affirmed, $p < .001$, $d = 1.26$ [0.88, 1.62], and affirmed, $p < .001$, $d = 0.91$ [0.54, 1.26], confrontations.

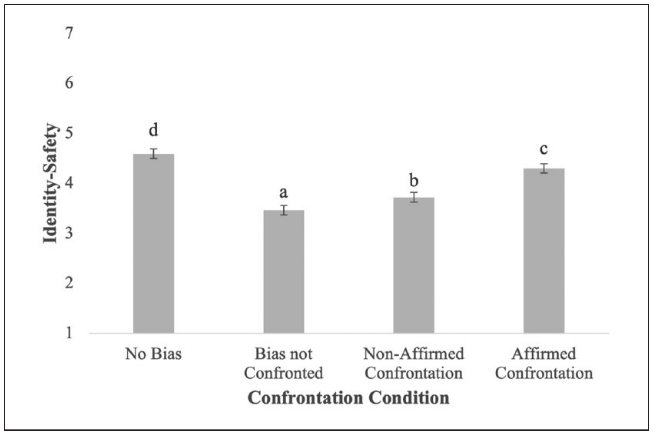
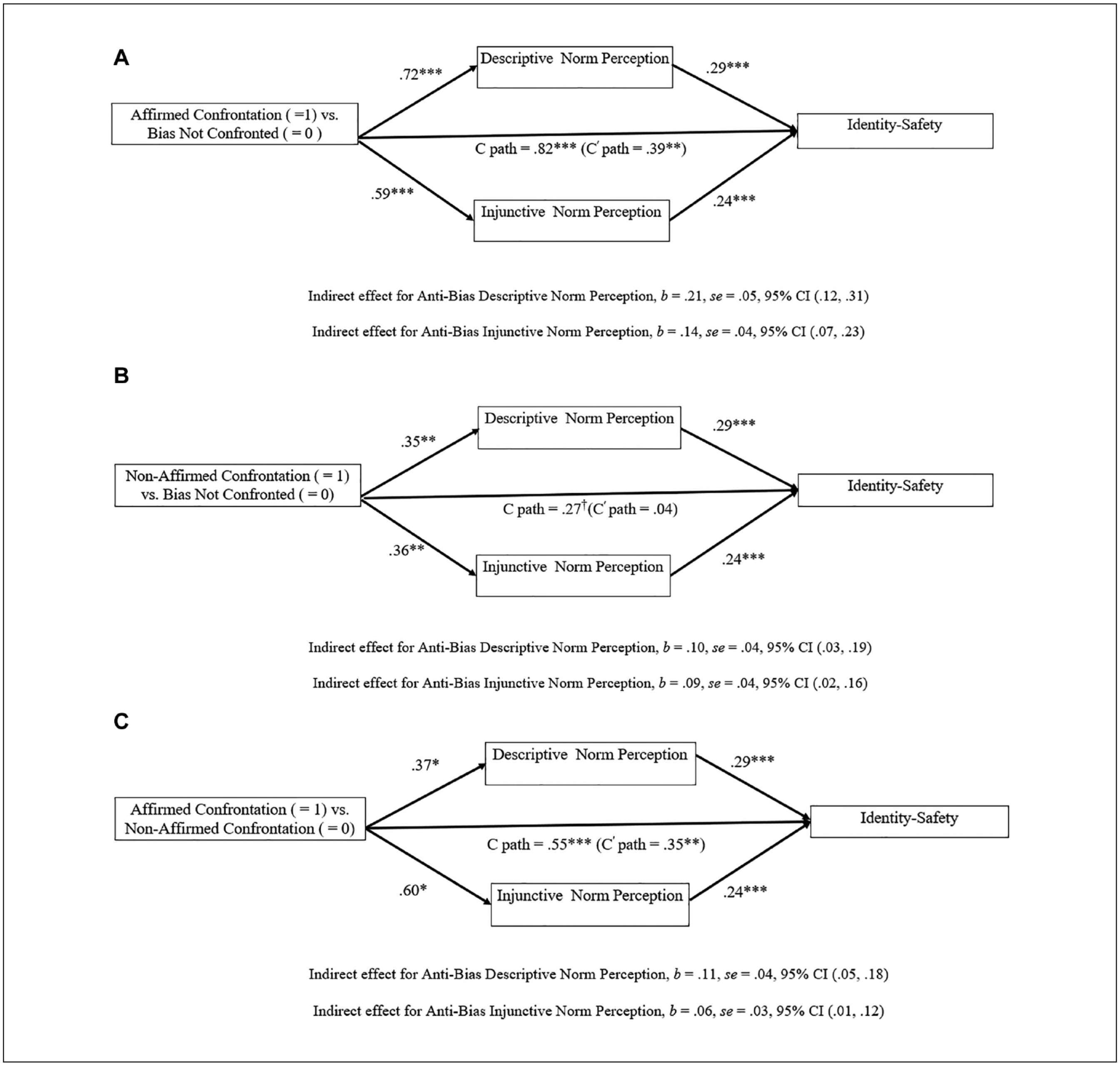


Figure 4. Identity-Safety as a Function of Confrontation Condition, Study 3.
Note. Error bars represent ±1 standard error. Means not sharing a lowercase letter differ significantly from each other.

Identity-Safety. Identity-safety was lower among Black ($M = 3.77$, $SD = 1.31$) than among White ($M = 4.28$, $SD = 1.09$) participants, $F(1, 554) = 30.59$, $p < .001$, $\eta_p^2 = .05$. A significant main effect of confrontation condition also emerged, $F(3, 554) = 30.61$, $p < .001$, $\eta_p^2 = .14$, but no interaction, $F(3, 554) = 1.83$, $p = .14$, $\eta_p^2 = .01$. As shown in Figure 4, identity-safety was substantially lower when bias occurred but was not confronted, relative to the no-bias condition, $p < .001$, $d = 1.02$ [0.76, 1.26]. Compared to leaving bias unconfronted, identity-safety was somewhat stronger with a non-affirmed confrontation, $p = .045$, $d = 0.22$ [−0.01, 0.46], and especially stronger with the affirmed confrontation, $p < .001$, $d = 0.70$ [0.46, 0.93]. Indeed, identity-safety was stronger in the affirmed than non-affirmed confrontation condition, $p < .001$, $d = 0.47$ [0.23, 0.70]. Nonetheless, identity-safety was still weaker in the affirmed, $p = .016$, $d = 0.29$ [0.06, 0.53], and non-affirmed, $p < .001$, $d = 0.78$ [0.53, 1.02], confrontation conditions than when bias did not occur in the first place.

Is the Effect of Confrontation on Identity-Safety Mediated by Norm Perception? We used Hayes' (2018) PROCESS macro


A
Affirmed Confrontation (=1) vs. Bias Not Confronted (= 0)
.72***
Descriptive Norm Perception
.29***
Identity-Safety
C path = .82*** (C′ path = .39**)
.59***
Injunctive Norm Perception
.24***
B
Non-Affirmed Confrontation (= 1) vs. Bias Not Confronted (= 0)
.35**
Descriptive Norm Perception
.29***
Identity-Safety
C path = .27†(C′ path = .04)
.36**
Injunctive Norm Perception
.24***
C
Affirmed Confrontation (= 1) vs. Non-Affirmed Confrontation (= 0)
.37*
Descriptive Norm Perception
.29***
Identity-Safety
C path = .55*** (C′ path = .35**)
.60*
Injunctive Norm Perception
.24***

Indirect effect for Anti-Bias Descriptive Norm Perception, *b* = .21, *se* = .05, 95% CI (.12, .31)

Indirect effect for Anti-Bias Injunctive Norm Perception, *b* = .14, *se* = .04, 95% CI (.07, .23)

Indirect effect for Anti-Bias Descriptive Norm Perception, *b* = .10, *se* = .04, 95% CI (.03, .19)

Indirect effect for Anti-Bias Injunctive Norm Perception, *b* = .09, *se* = .04, 95% CI (.02, .16)

Indirect effect for Anti-Bias Descriptive Norm Perception, *b* = .11, *se* = .04, 95% CI (.05, .18)

Indirect effect for Anti-Bias Injunctive Norm Perception, *b* = .06, *se* = .03, 95% CI (.01, .12)

Figure 5 Effects of Confrontation on Identity-Safety, Simultaneously Mediated by Anti-Bias Descriptive and Injunctive Norm Perception, Study 3

Note. Path values are standardized coefficients.

†p = .058. *$p < .05$. **$p < .01$. ***$p < .001$.

(Model 4; 5,000 bootstrapped samples) to test whether shifts in norms perceptions explained the differences in participants' identity-safety as a function of confrontation conditions. We excluded the no-bias condition. In a first analysis, the three conditions were coded to allow comparisons between (a) affirmed confrontation vs. bias not confronted and (b) non-affirmed confrontation vs. bias not confronted. We then modified the coding scheme to allow comparison between (c) the affirmed confrontation vs. non-affirmed confrontation. As shown in Figure 5, participants perceived stronger anti-bias descriptive and injunctive norms with confrontation than with bias not confronted, both when affirmed (Panel A) and not affirmed (Panel B), which in turn were associated with stronger identity-safety. Participants also perceived stronger descriptive and injunctive norms with the affirmed than non-affirmed confrontation, which in turn predicted identity-safety (Panel C). All three indirect effects were significant, as detailed in Figure 5.

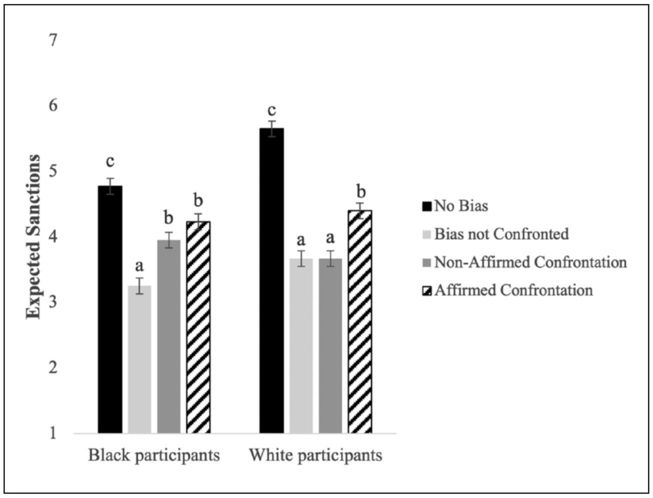


Figure 6 Expected Sanctions for Bias as a Function of Confrontation Condition, Study 3
Note. Error bars represent ±1 standard error. Within race, means not sharing a lowercase letter differ significantly from each other.

Expected Sanctions for Bias. As predicted, Black participants expected lower sanctions for being biased in the environment (M = 4.05, SD = 1.49) than White participants (M = 4.32, SD = 1.64), $F(1, 554) = 5.10$, $p = .011$, $\eta_p^2 = .01$. A significant confrontation condition main effect, $F(3, 554) = 39.53$, $p < .001$, $\eta_p^2 = .18$, was qualified by the interaction, $F(3, 554) = 4.20$, $p = .006$, $\eta_p^2 = .02$. As shown in Figure 6, Black and White participants showed the same patterns across confrontation conditions, but one unanticipated exception appeared responsible for the interaction.

Specifically, compared to the no-bias condition, both Black, $p < .001$, $d = 0.98$ [0.63, 1.32], and White, $p < .001$, $d = 1.48$ [1.09, 1.86], participants expected lower sanction when bias was not confronted. Among Black participants, both the non-affirmed, $p = .003$, $d = 0.49$ [0.15, 0.82], and affirmed, $p < .001$, $d = 0.70$ [0.37, 1.03], confrontations resulted in greater expected sanctions than when bias was not confronted. In contrast, we unexpectedly found that White participants rated sanctions equivalently in the non-affirmed confrontation and bias-not-confronted conditions, $p = 1.00$, although they reported greater sanctions when confrontation was affirmed than when bias was not confronted, $p = .003$, $d = 0.47$ [0.13, 0.81]. Given these patterns, we also observed that the non-affirmed and affirmed conditions were comparable among Black participants, $p = .233$, whereas White participants reported stronger expected sanctions in the affirmed than non-affirmed condition, $p = .003$, $d = 0.48$ [0.14, 0.82]. Finally, confrontation was associated with lower expected sanctions relative to the no-bias condition among both Black and White participants whether it was non-affirmed (Black: $p < .001$, $d = 0.57$ [0.23, 0.90]; White: $p < .001$, $d = 1.52$ [1.13, 1.90]) or affirmed (Black: $p = .025$, $d = 0.38$ [0.05, 0.69]; White: $p < .001$, $d = 0.91$ [0.54, 1.27]).

Future Intentions to Monitor Biases

A one-way ANOVA among White participants (i.e., those who completed this measure) indicated the confrontation condition main effect was not significant, $F(3, 259) = .20$, $p = .899$, $\eta_p^2 = .002$.

Social Costs. As in Study 2, the social costs index was formed by averaging the standardized scores on trait ratings of the confronter ($\alpha = .84$; $M = 3.00$, $SD = 1.30$) and evaluation/desire for future contact with the confronter ($\alpha = .92$; $M = 3.48$, $SD = 1.60$), reliability of the composite = .92. A significant race main effect, $F(1, 554) = 5.10$, $p = .024$, $\eta_p^2 = .01$, revealed that Black participants ($M = .05$, $SD = 0.93$) rated Dan more negatively than White participants ($M = -.11$, $SD = 0.79$), $d = 0.19$ [0.02, 0.35]. Replicating Study 2, the anticipated confrontation main effect was significant, $F(3, 554) = 11.34$, $p < .001$, $\eta_p^2 = .06$. The interaction between race and confrontation condition was not significant, $F(3, 554) = 0.81$, $p = .487$, $\eta_p^2 = .004$.

Examination of the confrontation main effect indicated that participants levied greater social costs toward Dan when he failed to confront bias ($M = 0.29$, $SD = 0.80$), relative to the no-bias condition ($M = 0.00$, $SD = 0.82$), $p = .004$, $d = 0.36$ [0.13, 0.60]. Participants also levied greater social costs toward Dan when he failed to confront bias relative to the non-affirmed ($M = -0.26$, $SD = 0.91$), $p < .001$, $d = 0.65$ [0.40, .88], and affirmed ($M = -0.16$, $SD = 0.85$), $p < .001$, $d = 0.54$ [0.31, 0.78], confrontations. In addition, participants levied fewer social costs toward Dan when he was a lone confronter than when no bias was expressed, $p = .012$, $d = 0.30$ [0.06, 0.53]. Finally, the no bias and non-affirmed confrontation conditions did not differ, $p = .116$, nor did the non-affirmed and affirmed confrontation conditions, $p = .341$. Overall, these results indicate that observers especially dislike a person who does not speak out against bias.

Discussion

Study 3 demonstrated the norm-signaling function of bias confrontation among Black as well as White participants. Extending research by Hildebrand et al. (2020), we found that these anti-bias norm perceptions, in turn, were associated with enhanced identity-safety compared to when bias went unconfronted. Similarly, confrontation especially boosted Black participants' belief that the company would sanction bias. Furthermore, replicating Study 2, observers levied social costs against a person who did not (vs. did) confront—an interesting result that we address in the General Discussion section.

A couple of unexpected effects also emerged. First, among White participants, the impact of non-affirmed confrontation on norm perceptions was weaker than that in our other studies, although patterns conformed to predictions. Nonetheless, an integrative data analysis (Curran & Hussong, 2009) across Studies 2 and 3 indicated that, compared to when bias not confronted, both the affirmed and non-affirmed confrontations significantly strengthened norm perceptions (for descriptive norms, ps ≤ .002; for injunctive norms, ps ≤ .001), and the non-affirmed vs. affirmed confrontation conditions did not differ from each other (for descriptive norms, p = .848; for injunctive norms, p = .100). Second, White participants' intentions to monitor their racial biases were unaffected by confrontation. We may have encountered a ceiling effect given strong social norms against bias toward Black people (e.g., Crandall et al., 2002) and the explicit nature of this measure.

General Discussion

Across three studies concerning two types of bias and for target- and non-target group members, bias confrontation communicated to observers that bias is neither common (anti-bias descriptive norms) nor accepted (anti-bias injunctive norms) in the local environment. Studies 1 and 2 suggested a restorative function of confrontation after bias occurs: Among non-target observers, bias confrontation strengthened the perception of anti-bias descriptive norms compared to leaving bias unconfronted and restored the perception of anti-bias injunctive norms to the baseline (i.e., when no bias had occurred). Study 3 demonstrated that the norm-signaling function of confrontation is applicable to anti-Black bias among both Black and White participants. Furthermore, replicating the work of Hildebrand et al. (2020), Black participants who observed a confrontation reported stronger identity-safety in the environment than when bias was not confronted, and especially when the confrontation was affirmed. We also observed this pattern; moreover, the positive effects of confrontation were statistically mediated by perceptions of anti-bias descriptive and injunctive norms. Together, these findings indicate that confrontation effectively signals anti-bias norms to observers and contributes to identity-safety in the face of bias.

The current research extends past research that has focused on confrontation outcomes for the person confronted and confronter (see Monteith et al., 2022) to consider confrontation's broader influence on perceptions of social norms. Past research has shown that anti-bias norms can be signaled in various ways (Moser & Branscombe, 2022; Murrar et al., 2020) but did not involve situations where bias had just occurred. Our research also extends prior work by distinguishing between descriptive and injunctive norms, rather than assessing norm perceptions without this distinction (c.f., Koudenburg et al., 2021). The norm literature suggests that injunctive norms have advantages over descriptive norms by increasing prosocial action even in settings characterized by antisocial action and by enhancing norm-congruent behavior in environments both similar to and different from those in which the norms are made salient (Cialdini et al., 1990). This paints a promising picture on the potential social control effects of bias confrontation on constraining bystanders' bias expressions (Kalkstein et al., 2023).

Our findings suggest that observers explain confrontation with dispositional more than situational attributions (Kelley, 1967). Contrary to the possibility of explaining confrontation in terms of negative confronter evaluations (e.g., complainer), Studies 2 and 3 revealed that confronting produced *lower* social costs than not confronting. Interestingly, we did not find that consensus information—conveyed with confrontation affirmations—boosted perceptions of anti-bias norms beyond levels observed with a lone confronter. This result runs contrary to Kelley's (1967) covariation model. Perhaps consensus information would need to be more extensive, reflecting the views of more people (see Wells & Harvey, 1977). Our materials communicated consensus information among a few people involved in a workplace conversation, whereas our norm measures referenced the entire company.

We found it interesting that participants evaluated the confronter more positively than a counterpart who did not confront, considering the robust literature indicating that confrontation has social costs (for a review, see Monteith et al., 2022). However, unlike past research, observers of confrontation were the focus in the current work, rather than people who were confronted themselves. According to recent theorizing about social costs (Monteith et al., 2022), they increase to the extent that the perceiver believes the confronter is trying to impugn their nonprejudiced self-image. Perhaps observers do not sense perceived impugnment, and this accounts for relatively positive evaluations of confronters.

The present research has clear practical implications. If people realize that confronting bias serves as a powerful tool to signal and encourage egalitarian normative climates, and that it fosters identity-safety for target group members, they may be more likely to engage in confrontation.

Limitations and Future Directions

Aspects of our methodology may have maximized confrontation's effects on social norm perceptions, and future research is critical for addressing possible boundary conditions. First, the confronter was a dominant group ally (i.e., White male), which may have minimized negative impressions (e.g., complainer, overly sensitive) that are commonly found to be greater when target group members confront (Gervais & Hillard, 2014; Schultz & Maddox, 2013). Whether observers make weaker social norm inferences with target group confronters is important to investigate. Studying how the group targeted by the bias perceives confrontations made by their ingroup is also important. On one hand, targets may understand their ingroup's influence as less powerful than the majority group (Droogendyk et al., 2016). On the other hand, ally confronters may raise targets' suspicion of the motives behind the confrontation (e.g., self-serving vs.

sincere motives; Burns & Granz, 2023; Chu, Ashburn-Nardo, 2022), and this suspicion may reduce targets' perceptions of anti-bias norms in the environment.

Second, the current research involved a blatant act of bias. Would confrontation of a subtly biased remark or stereotypic response likewise shape observers' perceptions of anti-bias social norms or result in confronter derogation instead? Might reactions to observing subtle bias confrontations be especially negative with a target group confronter?

Third, the confrontation language used in the current research captured key components of confrontations that are valued by Black people, including directly communicating disapproval and labeling the comment as prejudiced (Bak et al., 2023). These components may convey genuine and intrinsic concerns about the treatment of target group members, rather than the desire to accrue personal benefits (Chu & Ashburn-Nardo, 2022; Kutlaca & Radke, 2023; Radke et al., 2020). Future research is needed to understand how confrontation style affects the perception of social norms.

Finally, future research is needed to move beyond the current scenario-based method to determine whether results generalize to situations in which confrontation is actually observed.

Conclusion

People's behaviors are governed by norms (Kalkstein et al., 2023). Social norms are not fixed features of environments, nor are they beyond the influence of individual action. People have the capacity to construct norms and to influence others to encourage egalitarianism. Our research shows that bias confrontations offer this opportunity for individuals to reset ground rules in the face of prejudice. Additional research is needed to better understand the potential power and limitations of confrontation for shaping anti-bias norms. However, the current research offers an optimistic start by suggesting that interpersonal confrontations influence norm perceptions and help to foster identity-safety among individuals in situations where bias occurs.

Acknowledgments

The authors would like to thank the previous and current members of the Monteith Intergroup Relations lab for their helpful feedback and assistance.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Margo J. Monteith https://orcid.org/0000-0002-3427-9164

Supplemental Material

Supplemental material is available online with this article.

Notes

1. In all studies, α increases up to .90 if the first critical item is dropped. Results do not differ with the two-item index.
2. Our pre-registration also specified internal and external motivations to respond without prejudice (Plant & Devine, 1998) for exploratory purposes. No results involving motivation were significant, see SM.
3. Two additional attribution items were, in retrospect, oddly worded and did not correlate with the face-valid counterparts. These items were excluded.
4. Data collection for Black participants on MTurk moved much slower than for White participants. We therefore supplemented recruitment of Black participants (58%) with Prime Panels candidates.

References

Alt, N. P., Chaney, K. E., & Shih, M. J. (2019). "But that was meant to be a compliment!:" Evaluative costs of confronting positive racial stereotypes. *Group Processes and Intergroup Relations*, *22*(5), 655–672. https://doi.org/10.1177/1368430218756493

Ashburn-Nardo, L., Morris, K. A., & Goodwin, S. A. (2008). The confronting prejudiced responses (CPR) model: Applying CPR in organizations. *Academy of Management Learning & Education*, *7*(3), 332–342. https://doi.org/10.5465/amle.2008.34251671

Baer, J. S. (1994). Effects of college residence on perceived norms for alcohol consumption: An examination of the first year in college. *Psychology of Addictive Behaviors*, *8*, 43–50. https://doi.org/10.1037/0893-164X.8.1.43

Bak, H., Jurcevic, I., & Trawalter, S. (2023). What black people value when white people confront prejudice. *The Journal of Social Psychology*, *164*(2), 187–198. https://doi.org/10.1080/00224545.2023.2178875

Blanchard, F. A., Crandall, C. S., Brigham, J. C., & Vaughn, L. A. (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology*, *79*(6), 993–997. https://doi.org/10.1037/0021-9010.79.6.993

Blanton, H., Köblitz, A., & McCaul, K. D. (2008). Misperceptions about norm misperceptions: Descriptive, injunctive, and affective "social norming" efforts to change health behaviors. *Social and Personality Psychology Compass*, *2*, 1379–1399. https://doi.org/10.1111/j.1751-9004.2008.00107

Burns, M. D., & Granz, E. L. (2023). "Sincere White people, work in conjunction with us": Racial minorities' perceptions of White ally sincerity and perceptions of ally efforts. *Group Processes and Intergroup Relations*, *26*, 453–474. https://doi.org/10.1177/13684302211059699

Chaney, K. E., & Sanchez, D. T. (2018). The endurance of interpersonal confrontations as a prejudice reduction strategy. *Personality and Social Psychology Bulletin*, *44*(3), 418–429. https://doi.org/10.1177/0146167217741344

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology*, *97*(6), 1045–1060. https://doi.org/10.1037/a0016239

Chu, C., & Ashburn-Nardo, L. (2022). Black Americans' perspectives on ally confrontations of racial prejudice. *Journal of Experimental Social Psychology*, *101*, Article 104337. https://doi.org/10.5465/AMBPP.2022.15150abstract

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 151–192). McGraw-Hill.

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359–378. https://doi.org/10.1037/0022-3514.82.3.359

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

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

Droogendyk, L., Wright, S. C., Lubensky, M., & Louis, W. R. (2016). Acting in solidarity: Cross-group contact between disadvantaged group members and advantaged group allies. *Journal of Social Issues*, *72*(2), 315–334. https://doi.org/10.1111/josi.12168

Drury, B. J., & Kaiser, C. R. (2014). Allies against sexism: The role of men in confronting sexism. *Journal of Social Issues*, *70*(4), 637–652. https://doi.org/10.1111/josi.12083

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Gervais, S. J., & Hillard, A. L. (2014). Confronting sexism as persuasion: Effects of a confrontation's recipient, source, message, and context. *Journal of Social Issues*, *70*(4), 653–667. https://doi.org/10.1111/josi.12084

Giner-Sorolla, R. (2018, January 24). *Powering your interaction* [Blog post]. https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2

Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*(3), 491–512. https://doi.org/10.1037/0022-3514.70.3.491

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Hildebrand, L. K., Jusuf, C. C., & Monteith, M. J. (2020). Ally confrontations as identity-safety cues for marginalized individuals. *European Journal of Social Psychology*, *50*, 1318–1333. https://doi.org/10.1002/ejsp.2692

Hildebrand, L. K., Monteith, M. J., & Arriaga, X. B. (2023). The role of trust in reducing confrontation-related social costs. *Journal of Personality and Social Psychology*. Advanced online publication. https://doi.org/10.1037/pspi0000429

Johnson, I. R., & Pietri, E. S. (2023). Ally endorsement: Exploring allyship cues to promote perceptions of allyship and positive STEM beliefs among White female students. *Group Processes & Intergroup Relations*, *26*(3), 738–761. https://doi.org/10.1177/13684302221080467

Kalkstein, D. A., Hook, C. J., Hard, B. M., & Walton, G. M. (2023). Social norms govern what behaviors come to mind—And what do not. *Journal of Personality and Social Psychology*, *124*(6), 1203–1229. https://doi.org/10.1037/pspi0000412

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, *15*, 192–238.

Koudenburg, N., Kannegieter, A., Postmes, T., & Kashima, Y. (2021). The subtle spreading of sexist norms. *Group Processes & Intergroup Relations*, *24*(8), 1467–1485. https://doi.org/10.1177/1368430220961838

Kutlaca, M., Becker, J., & Radke, H. (2020). A hero for the outgroup, a black sheep for the ingroup: Societal perceptions of those who confront discrimination. *Journal of Experimental Social Psychology*, *88*, Article 103832. https://doi.org/10.1016/j.jesp.2019.103832

Kutlaca, M., & Radke, H. R. (2023). Towards an understanding of performative allyship: Definition, antecedents and consequences. *Social and Personality Psychology Compass*, *17*(2), Article e12724. https://doi.org/10.1111/spc3.12724

Major, B., & O'Brien, L. T. (2005). The social psychology of stigma. *Annual Review of Psychology*, *56*, 393–421.

Major, B., Quinton, W. J., & Schmader, T. (2003). Attributions to discrimination and self-esteem: Impact of group identification and situational ambiguity. *Journal of Experimental Social Psychology*, *39*(3), 220–231. https://doi.org/10.1016/S0022-1031(02)00547-4

Mallett, R. K., & Wagner, D. E. (2011). The unexpectedly positive consequences of confronting sexism. *Journal of Experimental Social Psychology*, *47*(1), 215–220. https://doi.org/10.1016/j.jesp.2010.10.001

Miller, D. T., & Prentice, D. A. (2016). Changing norms to change behavior. *Annual Review of Psychology*, *67*, 339–361. https://doi.org/10.1146/annurev-psych-010814-015013

Monteith, M. J., Deneen, N. E., & Tooman, G. D. (1996). The effect of social norm activation on the expression of opinions concerning gay men and Blacks. *Basic and Applied Social Psychology*, *18*(3), 267–288. https://doi.org/10.1207/s15324834basp1803_2

Monteith, M. J., Mallett, R. K., & Hildebrand, L. K. (2022). Confronting intergroup biases: Validity and impugnment as determinants of other-confrontation consequences. *Advances in Experimental Social Psychology*, *66*, 1–57. https://doi.org/10.1016/bs.aesp.2022.04.001

Moser, C. E., & Branscombe, N. R. (2022). Male allies at work: Gender-equality supportive men reduce negative underrepresentation effects among women. *Social Psychological and Personality Science*, *13*(2), 372–381. https://doi.org/10.1177/19485506211033748

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science*, *18*(10), 879–885. https://doi.org/10.1111/j.1467-9280.2007.01995.x

Murrar, S., Campbell, M. R., & Brauer, M. (2020). Exposure to peers' pro-diversity attitudes increases inclusion and reduces the achievement gap. *Nature Human Behavior*, *4*(9), 889–897. https://doi.org/10.1038/s41562-020-0899-5

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, *113*(3), 566–571. https://doi.org/10.1073/pnas.1514483113

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

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *753*(3), 811–832. https://doi.org/10.1037/0022-3514.75.3.811

Purdie-Vaughns, V., Steele, C. M., Davies, P. G., Ditlmann, R., & Crosby, J. R. (2008). Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology*, *94*(4), 615–630. https://doi.org/10.1037/0022-3514.94.4.615

Radke, H. R., Kutlaca, M., Siem, B., Wright, S. C., & Becker, J. C. (2020). Beyond allyship: Motivations for advantaged group members to engage in action for disadvantaged groups. *Personality and Social Psychology Review*, *24*(4), 291–315. https://doi.org/10.1177/1088868320918698

Rattan, A., Kroeper, K., Arnett, R., Brown, X., & Murphy, M. (2023). Not such a complainer anymore: Confrontation that signals a growth mindset can attenuate backlash. *Journal of Personality and Social Psychology*, *124*(2), 344–361. https://doi.org/10.1037/pspi0000399

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104–112. https://doi.org/10.1037/0022-3514.64.1.104

Schultz, J. R., & Maddox, K. B. (2013). Shooting the messenger to spite the message? Exploring reactions to claims of racial bias. *Personality and Social Psychology Bulletin*, *39*(3), 346–358. https://doi.org/10.1177/0146167212475223

Shelton, J. N., & Stewart, R. E. (2004). Confronting perpetrators of prejudice: The inhibitory effects of social costs. *Psychology of Women Quarterly*, *28*(3), 215–223. https://doi.org/10.1111/j.1471-6402.2004.00138.x

Stangor, C., Sechrist, G. B., & Jost, J. T. (2001). Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin*, *27*(4), 486–496. https://doi.org/10.1177/0146167201274009

Sue, D. W., & Spanierman, L. (2020). *Microaggressions in everyday life*. John Wiley & Sons.

Swim, J. K., & Hyers, L. L. (1999). Excuse me—What did you just say?!: Women's public and private responses to sexist remarks. *Journal of Experimental Social Psychology*, *35*(1), 68–88. https://doi.org/10.1006/jesp.1998.1370

Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, *10*(1), 181–211. https://doi.org/10.1111/sipr.12022

Wells, G. L., & Harvey, J. H. (1977). Do people use consensus information in making causal attributions? *Journal of Personality and Social Psychology*, *35*, 279–293. https://doi.org/10.1037/0022-3514.35.5.279

Wilton, L. S., Rattan, A., & Sanchez, D. T. (2018). White's perceptions of biracial individuals' race shift when biracials speak out against bias. *Social Psychological and Personality Science*, *9*, 887–1019. https://doi.org/10.1177/1948550617731497