Group Processes & Intergroup Relations
G P I R

Article

Social norms of prejudice confrontations impact anticipated costs and benefits of confronting prejudice

Group Processes & Intergroup Relations
1–27

DOI: 10.1177/13684302251355900
journals.sagepub.com/home/gpi
Sage

Izilda Pereira-Jorge and Kimberly E. Chaney

Abstract

The current research examined the impact of perceived social norms of confronting interpersonal prejudice on White people's perceived barriers and intentions to confront prejudice. The present research (N_{total} = 1,057) manipulated injunctive (Study 1) and descriptive (Studies 1–3) social norms of confronting prejudice via group consensus information and group behavior. Across studies, descriptive confronting norms facilitated perceptions of fewer anticipated social costs (Studies 1, 3) and greater social benefits (Studies 1–3) relative to the absence of a confronting norm (Studies 2–3) or an injunctive norm of confronting prejudice (Study 1). While social norms of prejudice confrontation did not directly impact White people's intentions to confront prejudice (Study 1–3) or a specific instance of anti-Asian bias (Studies 2, 3), social norms did indirectly elicit greater intentions to confront prejudice through expectations of reduced social costs and increased social benefits (Studies 1, 2, 3).

Keywords

anti-Asian bias, confronting prejudice, social benefits, social costs, social norms

Paper received 1 October 2024; revised version accepted 30 May 2025.

Interpersonal prejudice confrontations have been a successful tool in reducing White people's stereotype use in the moment (Czopp et al., 2006), a week after confrontation (Chaney & Sanchez, 2018), and up to a month after (e.g., Chaney et al., 2025; Munger, 2017). Prejudice confrontations are defined as verbal or nonverbal expressions of disapproval of an individual's display of bias (Chaney & Chasteen, 2023; Shelton et al., 2006). Though prejudicial remarks remain prevalent in physical and virtual spaces, only 30–50% of people confront prejudice (Dickter & Newton, 2013; Hurd et al., 2022). Primary barriers to confronting prejudice are a belief that confrontation will not reduce a perpetrator's bias (i.e., low perceived interpersonal social benefits for the confronter; Good et al., 2012; Kaiser & Miller, 2004; Rattan & Dweck, 2010) and a fear of how the

University at Buffalo, USA

Corresponding author:
Izilda Pereira-Jorge, University at Buffalo, SUNY, 204 Park Hall, North Campus, Buffalo, NY 14260-1660, USA.
Email: izildape@buffalo.edu

perpetrator and others will react to a confrontation (i.e., high perceived interpersonal social costs for the confronter; Ashburn-Nardo et al., 2014; Czopp, 2019; Good et al., 2012).

The current research examined one way to mitigate these barriers: social norm messaging. Social norm messaging, that is, communicating a social norm, has been found to elicit attitudes and behaviors in line with the norm (e.g., egalitarian attitude social norm messaging facilitates greater egalitarian attitudes; Blanchard et al., 1994; Murrar et al., 2020). Perceived social norms have been proposed as a critical component in eliciting desired behaviors (e.g., increasing recycling; Viscusi et al., 2011) or inhibiting undesired behaviors (e.g., reducing drinking in college campuses; Perkins & Craig, 2002). Thus, in three experimental studies, we examine if social norm messaging about prejudice confrontations would facilitate greater prejudice confrontation intentions among White Americans. Perceived social costs and benefits of confronting prejudice were examined as mechanisms.

Barriers to Prejudice Confrontation

While people indicate they are highly likely to confront prejudice in hypothetical scenarios (Hurd et al., 2022; Swim & Hyers, 1999; Woodzicka & LaFrance, 2001), few people confront prejudice in real scenarios (Ashburn-Nardo et al., 2014; Dickter, 2012; Hyers, 2007; Kawakami et al., 2009; Swim & Hyers, 1999; Woodzicka & LaFrance, 2001). Identifying methods to promote greater rates of prejudice confrontations is thus a critical step in mitigating bias. Even when someone has detected an instance of prejudice, feels it is urgent to address, and feels personally responsible for addressing it, the Confronting Prejudiced Responses (CPR) model (Ashburn-Nardo & Karim, 2019; Ashburn-Nardo et al., 2008) predicts that people are still unlikely to confront prejudice when they anticipate social costs that outweigh anticipated social benefits. Supporting this theoretical model, research on prejudice confrontations has highlighted that perceived social costs and benefits serve as mechanisms that predict lower rates of prejudice confrontation (Ashburn-Nardo et al., 2014; Good et al., 2012; Hyers, 2007; Rattan & Dweck, 2010; Sechrist, 2010).

Definitions of perceived social benefits of confronting prejudice primarily focus on curbing a perpetrator's future bias (Good et al., 2012; Hyers, 2007). Indeed, one's belief that prejudice is malleable and can be decreased predicts greater confrontation intentions (Chaney & Chasteen, 2023; Rattan & Dweck, 2010). Other social benefits of prejudice confrontations include signaling to marginalized groups that their marginalized identities are valued in that environment (Chu & Ashburn-Nardo, 2022; Hildebrand et al., 2020; Li et al., 2024).

Social costs of confronting prejudice center on potential backlash from the perpetrator of prejudice or others who witnessed the confrontation (e.g., being perceived as hypersensitive, a troublemaker; Czopp et al., 2006; Good et al., 2012; Kaiser & Miller, 2001, 2003, 2004; Wessel et al., 2023). Concerns about the social costs of confronting prejudice are not unfounded. White people who have been confronted for a prejudiced comment do at times react in a retaliatory way toward the confronter (Wessel et al., 2023), including expressing hostility and denial of their bias (e.g., Czopp et al., 2006; Wessel et al., 2023). When confronted by a Black person rather than a White person, White perpetrators of anti-Black bias perceive the confrontation as less legitimate and ruder (Czopp & Monteith, 2003; Rasinski & Czopp, 2010). Nevertheless, anticipated social costs are a factor in both advantaged and marginalized group member's decisions to confront prejudice (Ashburn-Nardo & Karim, 2019; Czopp, 2019).

As perceived social costs and benefits of prejudice confrontations have been identified as critical mechanisms for promoting it, identifying strategies to increase White people's perceived benefits and/or to decrease perceived costs of confronting prejudice is imperative to promote a behavior that both reduce perpetrator bias (Chaney & Sanchez, 2018) and promote feelings of inclusion for marginalized group members (Hildebrand et al., 2020; Li et al., 2024). We

propose that signaling that prejudice confrontation is a norm-adhering behavior may reduce perceived costs and increase perceived benefits.

Social Norms as a Vehicle to Facilitate Egalitarian Behavior

Individuals may align their attitudes and behaviors to fit perceived social norms conveyed in one's local environment (e.g., Cialdini et al., 1990; Cialdini et al., 2006) or explicitly sanctioned by one's institutions (e.g., enacting gay marriage laws in the US resulted in more favorable attitudes toward sexual minorities; Ofosu et al., 2019; Tankard & Paluck, 2017). Social norms may be signaled through various sources of information (i.e., social norm messaging), such as group consensus information, observed group members' behavior, and institutional influences (Tankard & Paluck, 2016). Moreover, perceived social norms may be descriptive or injunctive; descriptive norms indicate that the majority of one's group performs the normed behavior, while injunctive norms indicate that a behavior is approved by others and is expected of oneself (Cialdini et al., 1990, 2006; Cialdini & Trost, 1998).

Critically, *social norm messaging* has been identified as a tool to change intergroup attitudes through both group consensus information and observed ingroup members' behaviors. For example, overhearing someone condemn racism or express pro-Black attitudes, compared to hearing anti-Black opinions, reduces expression of racist opinions—demonstrating the impact of local egalitarian norms (i.e., observed ingroup member's behavior) on curbing bias expression in that moment (Blanchard et al., 1994; Monteith et al., 1996) and up to 1 month later (Munger, 2017; Zitek & Hebl, 2007). Broader egalitarian social norm messaging with classroom posters or videos (i.e., group consensus information) that explicitly stated that the majority of college students hold egalitarian attitudes, compared to the absence of such messaging, increased White students' rejection of discrimination and appreciation of diversity (Murrar et al., 2020). Further, these positive intergroup attitude outcomes were mediated by perceived egalitarian norms amongst peers, rather than perceived university commitment to diversity (Murrar et al., 2020), suggesting that perceived norms centering egalitarian behaviors are critical in cultivating inclusive attitudes and behaviors.

More directly related to prejudice confrontation, when one or two individuals affirmed a prejudice confrontation, the environment was perceived as more egalitarian than when a prejudice confrontation was not affirmed (Li et al., 2024). Further, strong norms promoting confrontation behavior (e.g., all bystanders confront racist hate speech), compared to weaker norms (e.g., one of the bystanders confronts), are critical for reducing perceptions of harm from the hate speech (Zapata et al., 2024). This may indicate that confrontations may be perceived as mitigating the harm caused by hate speech only when such confronting norms are strongly adhered to (Zapata et al., 2024). These findings suggest that prejudice confrontation behaviors may signal a norm of egalitarianism, though assessment of a norm of confronting prejudice as a vehicle to promote more prejudice confrontations was not made. We propose that signaling a social norm of confronting prejudice may promote greater prejudice confrontation behavior.

The Current Research

The role of egalitarian norms has been theorized as a missing component of the CPR model that considers how an individual's broader environment may signal prejudice confrontation as an acceptable, low-cost behavior (De Souza & Schmader, 2022; Nelson et al., 2011). Shifting beyond a norm of egalitarian attitudes to a norm of egalitarian behavior, such as prejudice confrontations, affords people a concrete way to enact egalitarian ideologies. The present research seeks to examine if social norm messaging can be harnessed to promote lower perceived costs and greater perceived benefits of confronting prejudice, resulting in greater prejudice confrontation intentions. Normative behaviors are perceived as socially acceptable, and thus permissible to engage in due in part to heightened expected

social benefits for performing the behavior (Rimal et al., 2005). As such, we propose that social norm messaging of prejudice confrontations will reduce anticipated interpersonal costs and boost anticipated interpersonal benefits to individuals contemplating addressing a prejudicial comment, relative to the absence of such norms. Specifically, a social norm of confronting prejudice should signal that confronting is not a costly behavior, but rather an appropriate and positive one that is in line with people's expectations of acceptable behavior. Further, a social norm of confronting prejudice may increase perceived benefits, as such a norm should suggest confrontation is a useful and effective strategy for dealing with prejudiced comments or behaviors. We focus on White Americans who may be ideal confronters of racism due to their high societal power (e.g., Sidanius & Pratto, 2001). Further, White Americans' attributions to prejudice are perceived as more legitimate compared to marginalized group members' (Rasinski & Czopp, 2010; Schultz & Maddox, 2013), and thus they may be more likely to promote bias reduction in perpetrators.

In three studies, we examine how prejudice confrontation social norm type (i.e., descriptive vs. injunctive) and information source (i.e., group consensus information, observed behavior) impact White Americans' perceived social costs and benefits of others and themselves confronting prejudice, with implications for their intentions to confront prejudice. When examining social norm messaging via observed behavior, we also examine identity-absent messaging of confronting social norms (Studies 1–2) and ingroup referents (Study 3).

The primary hypotheses are outlined below. Study 3 was preregistered on the Open Science Framework (OSF; https://osf.io/6g7yp/?=61be4d8fac824dcabb576a3935cc8d78). All materials and data for all studies are available at the OSF as well (https://osf.io/v7kwz/?=4f6412282a644105b3c808d4b6c563ed). We report all manipulations, measures, and exclusions across all studies. All studies were conducted with Institutional Review Board (IRB) approval at the University of Connecticut, and all participants provided consent.

Hypothesis 1: Prejudice confrontation social norms (i.e., descriptive and injunctive social norms) will facilitate fewer anticipated social costs and greater anticipated social benefits of confronting prejudice compared to the absence of a social norm.

Hypothesis 2: A prejudice confrontation social norm, relative to the absence of such a norm, will elicit greater intentions to confront prejudice.

Hypothesis 3: Prejudice confrontation social norms, relative to the absence of a norm, will elicit greater intentions to confront prejudice through fewer anticipated social costs and greater anticipated social benefits of confronting prejudice (i.e., significant indirect effects in mediation analysis).

Study 1

Study 1 examined the impact of descriptive or injunctive prejudice confrontation social norms (e.g., Cialdini et al., 1990, 2006), compared to a no-information control condition, on White Americans' anticipated social costs and benefits of confronting prejudice. Social norms were manipulated via an ostensible news article reporting on scientific findings that explicitly communicated social norms through group consensus information (e.g., Tankard & Paluck, 2016).

Method

Participants. An a priori power analysis in G*Power (Faul et al., 2007) revealed a desired sample size of 303 participants to detect a small–medium effect ($d = 0.36$) with 80% power for a three-cell, between-subjects analysis of variance (ANOVA). In case of exclusions, 388 participants who identified as non-Hispanic White residing in the US were recruited from Prolific during February 2023 in exchange for compensation. Participants who did not identify as

non-Hispanic White ($n = 55$) and/or failed two or more attention checks ($n = 37$) were excluded from analyses. This left a final analytic sample of 333. See Table 1 for participant demographics for Studies 1–3.

Procedures. Study 1 was an online survey about perceptions of social interactions. Participants were randomly assigned to one of three conditions where they read an excerpt of an ostensible news article that described scientific findings (e.g., Williams & Eberhardt, 2008). In the Descriptive Norm condition, it was reported that 85% of Americans speak out against prejudice and discrimination when observing friends, family, or strangers act in a prejudicial manner (see Figure 1). In the Injunctive Norm condition, 85% of Americans reported thinking that people should speak out against prejudice when family, friends, or strangers act in a prejudicial manner. Finally, in the Control condition, social norms in a nonprejudice domain were presented (i.e., norms for using utensils when eating food). For full materials, see the Supplemental Material.

After reviewing this information, participants completed three manipulation check questions (e.g., "What was the news outlet of the article you just read?"). Participants who failed the third, critical manipulation check question ($n = 25$; "The main point of the article was. . .") were provided the condition materials again before completing the manipulation checks a second time. Next, participants completed measures of anticipated social costs and social benefits of confronting prejudice, and general confrontation intentions (for additional findings on exploratory measures not reported here, see the Supplemental Material).[1] Participants then reported demographics before being debriefed and compensated.

Materials

Anticipated social costs and social benefits of confronting. Participants were asked to imagine they had "confronted (i.e., indicated verbal disapproval) towards someone who displayed a prejudicial comment or action..." Then, participants indicated how likely the confronted individual and others who witnessed the confrontation would be to act negatively toward them (17 items, anticipated social costs; $\alpha = .86$) or positively towards them (17 items, anticipated social benefits; $\alpha = .86$) on a 7-point scale (1= *Not at all likely*, 7 = *Very likely*). Items were adapted from Good et al. (2012).

General confrontation intentions. Participants answered three items (developed by the authors of the present study) indicating how likely they would be to confront prejudice the next time they saw it. Answers were given on a 7-point scale (1 = *Not at all likely*, 7 = *Very likely*; e.g., "I would confront anyone who makes prejudicial comments in the future"; $\alpha = .89$).

Results

Analyses were conducted as one-way ANOVAs. Significant main effects were examined with Fisher's least significant difference (LSD) post hoc tests (see Table 2 for ANOVA results and descriptive statistics by condition).

Anticipated social costs and benefits of confronting. There was a significant main effect of condition on anticipated social costs of confronting prejudice. Participants anticipated fewer social costs of confronting prejudice in the Descriptive Norm condition than in the Injunctive Norm, $p = .012$, $d = 0.35$, 95% $CI_{meandiff}$[−0.46, −0.06], and Control conditions, $p < .001$, $d = 0.45$, 95% $CI_{meandiff}$[−0.55, −0.15]. Participants reported similar expectations of social costs in both the Injunctive Norm and Control conditions, $p = .383$, $d = 0.12$, 95% $CI_{meandiff}$[−0.29, 0.11].

There was a significant main effect of condition on anticipated social benefits of confronting prejudice. While participants anticipated similar social benefits of confronting prejudice in both the Descriptive Norm and Injunctive Norm conditions, $p = .148$, $d = 0.20$, 95% $CI_{meandiff}$[−0.05, 0.33], greater social benefits were perceived in the Descriptive Norm than in the Control condition, $p = .003$, $d = 0.41$, 95% $CI_{meandiff}$[0.10, 0.48]. Participants reported similar expectations of

Table 1. Demographic summary.

	Study 1 *n* (%)	Study 2 *n* (%)	Study 3 *n* (%)
Gender			
Women (cisgender and transgender)	167 (50.1)	145 (52.9)	242 (54.2)
Men (cisgender and transgender)	154 (48.2)	117 (42.7)	187 (41.5)
Nonbinary/genderqueer	3 (0.9)	7 (2.6)	15 (3.3)
Transgender identifying	6 (1.8)	7 (2.5)	5 (1.5)
Questions/don't know	3 (0.9)	1 (0.5)	2 (0.4)
Self-identified	2 (0.6)	4 (1.5)	4 (0.9)
Sexual orientation			
Lesbian/gay	12 (3.6)	13 (4.7)	16 (3.6)
Bisexual	25 (7.5)	27 (9.9)	48 (10.7)
Pansexual	5 (1.5)	8 (2.9)	11 (2.4)
Queer	4 (1.2)	4 (1.5)	2 (0.4)
Questioning/not sure	5 (1.5)	1 (0.4)	1 (0.2)
Asexual	3 (0.9)	6 (2.2)	4 (0.9)
Heterosexual	277 (83.2)	214 (78.1)	365 (81.1)
Self-identified	2 (0.6)	1 (0.4)	3 (0.7)
	Study 1 *M* (*SD*)	Study 2 *M* (*SD*)	Study 3 *M* (*SD*)
Age (in years)	43.52 (14.94)	41.23 (13.92)	45.7 (14.81)
Political orientation (1 = s*trongly conservative*, 7 = *strongly liberal*)	4.85 (1.81)	4.72 (1.90)	4.75 (1.86)

social benefits in both the Injunctive Norm and Control conditions, $p = .119$, $d = 0.20$, 95% $CI_{meandiff}[-0.04, 0.34]$.

General confrontation intentions. There was no significant effect of condition on general intentions to confront prejudice. Participants reported moderate intentions to confront prejudice across conditions.

Mediations. Parallel multicategorical mediation analyses were conducted examining if anticipated social costs and social benefits of confronting prejudice mediated the effect of social norm condition on general confrontation intentions: Contrast 1: Descriptive (0) versus Injunctive (1); Contrast 2: Descriptive (0) versus Control (1). Mediation analyses were conducted in PROCESS Version 4.2 (Hayes, 2018), employing 5,000 bootstrap samples (see Figure 2).

Contrast 1 did not yield significant indirect effects on general confrontation intentions via anticipated social costs, $B = 0.03$ ($SE = 0.04$), 95% $CI_{boot}[-0.02, 0.12]$, or anticipated social benefits, $B = -0.10$ ($SE = 0.07$), 95% $CI_{boot}[-0.24, 0.03]$. Contrast 2 revealed no significant indirect effect of condition on confrontation intentions via social costs, $B = 0.04$ ($SE = 0.05$), 95% $CI_{boot}[-0.03, 0.15]$, though a significant indirect effect via social benefits emerged, $B = -0.10$ ($SE = 0.04$), 95% $CI_{boot}[-0.20, -0.03]$. Relative to the Control condition, the Descriptive Norm condition elicited greater expectations of social benefits of confronting prejudice, and in turn, greater intentions to confront prejudice.

Discussion

Study 1 demonstrated that only descriptive social norms of confronting prejudice, compared to no

Figure 1. The ostensible news article in the Study 1 Descriptive Norm condition.


The New York Times

Give this article 24

New study finds 85% of Americans actually speak out against prejudice when they see it

By Webster Hargreeves | August 17, 2022 at 8:07 p.m. ES

Topics pertaining to prejudice and discrimination have long generated considerable public dialogue and scholarly work. Of particular interest to researcher Dr. Desmond is the question of how everyday people respond to such incidents. How do most people react when they witness prejudice and discrimination?

Dr. Desmond's recent research indicates that a vast majority of Americans speak out against prejudice and discrimination when they observe others expressing bias or engaging in differential treatment based on social categories. In a series of studies recently published in the academic journal *Psychological Science*, principal investigator Dr. Desmond found that most people verbalize disapproval when family, friends, or strangers act in a prejudicial manner. Of 1200 people surveyed over a six-month period, approximately 85% who witnessed prejudice (e.g., insulting jokes or language, stereotyping, or excluding others due to their social group membership) had verbally communicated their disagreement or disapproval to the perpetrator.

In sum, this large-scale research suggests that standing up in opposition to prejudice and discrimination is a common socially acceptable behavior. That is, most people actively oppose bias when they detect it. These findings replicate prior research from a multi-site investigation conducted by a separate research team. According to Dr. Desmond, this research helps to clarify societal norms about confrontation of prejudice and discrimination.

norm, predicted greater anticipated social benefits of confronting prejudice. In contrast, injunctive prejudice confrontation norms resulted in similar expectations of social benefits as the control condition, which was absent of confronting norm information. Further, descriptive prejudice confrontation norms facilitated fewer anticipated social costs compared to injunctive social norms, demonstrating that the *type* of social norm may be key to ameliorating White people's perceived social costs of prejudice confrontation. Only an indirect effect of descriptive norm messaging (compared to the control) on confrontation intentions via social benefits emerged. Because descriptive norms were related to both fewer anticipated costs and greater anticipated benefits relative to the control, we opted to focus solely on descriptive norms in Study 2.

Study 2 aimed to advance Study 1 in two ways. First, we aimed to determine if descriptive confrontation norms would again mitigate anticipated social costs and increase perceived benefits of confronting when manipulating the social norm via observed group behavior (e.g., Cialdini et al., 1990; Zapata et al., 2024) instead of group consensus information via ostensible research findings (e.g., Tankard & Paluck, 2016). Second, past research has demonstrated that the strength of a prejudice confrontation norm can shift perceptions of biased behavior. When U.K. individuals encountered a strong descriptive confrontation norm, such that all bystanders confronted a perpetrator's hate speech toward racial minorities, this unanimous condemnation reduced perceptions of the harm caused by hate speech compared to when a single bystander confronted or when no bystanders confronted (i.e., using photo-based vignettes; Zapata et al., 2024). Thus, Study 2 sought to investigate how the perceived strength of the descriptive confrontation norm (e.g., all confront, some confront, or none confront) may facilitate intentions to confront and mitigate barriers toward confrontation.

Study 2

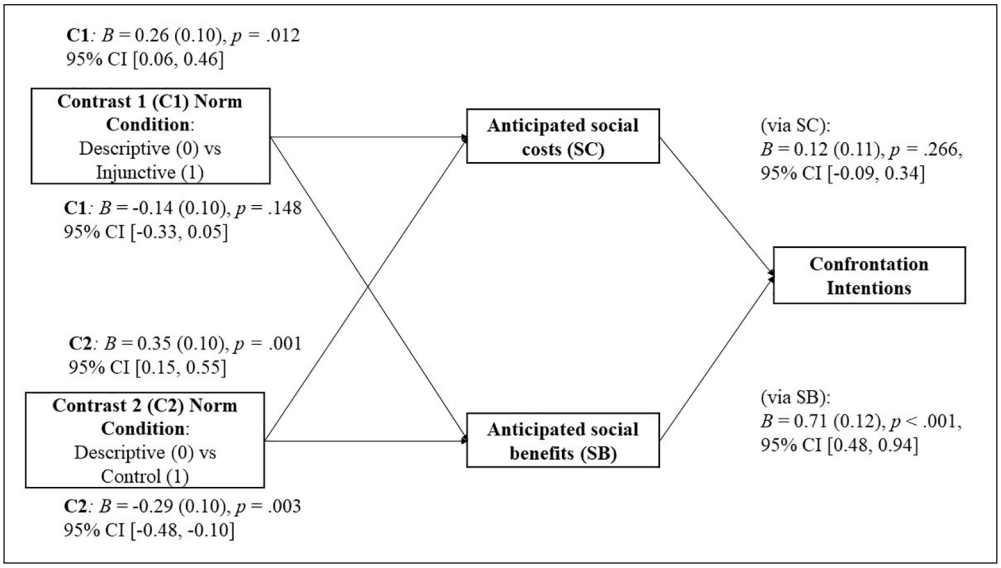
Study 2 aimed to replicate Study 1 findings with a new methodological manipulation of social norms and an examination of a specific kind of prejudice understudied in psychological research, anti-Asian racism (e.g., Alt et al., 2019; Meyers et al., 2020). Racism is prevalent in online spaces, with downstream consequences for the well-being of racial minorities, including Asian

Table 2. ANOVA results: Study 1.

	Condition effect			Descriptive Norm	Injunctive Norm	Control
Outcome	*F*(2, 330)	*p*	*d*	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)
Social costs	6.23	.002	0.39	4.18 (0.07)	4.44 (0.07)	4.53 (0.07)
Social benefits	4.50	.012	0.33	3.79 (0.06)	3.65 (0.07)	3.50 (0.07)
Confrontation intentions	1.17	.311	0.17	4.43 (0.14)	4.21 (0.15)	4.12 (0.14)

Note. ANOVA = analysis of variance.

Figure 2. Mediation: Study 1.



Note. Unstandardized regression coefficients are reported. Standard errors are reported in parentheses.

Americans (Hurd et al., 2022; M. H. J. Lee et al., 2024; R. T. Lee et al., 2019), highlighting the need to investigate confrontation in online spaces where anti-Asian bias may run rampant.

Following research on online confrontations (Hurd et al., 2022; Meyers et al., 2020), we manipulated descriptive social norms with observed confrontations from multiple social media users (e.g., Zapata et al., 2024). Participants were randomly assigned to see either no one confront a specific incident of anti-Asian prejudice (i.e., No Confronting Norm, 0%), one of six people confront (i.e., No Confronting Norm, 17%), or six of six people confront (i.e., Confronting Norm, 100%). Note that a condition in which one person confronts was included so that our manipulation specifically targeted social norms, not mere salience of the behavior. The present paradigm utilized raceless and genderless online personas who made the prejudiced remark and were either confronted or not confronted by other raceless and genderless online personas. In doing so, we aimed to isolate the effects of norms from the role of perpetrator and (non)confronter identities.

Hypotheses for anticipated social costs, benefits, and intentions to confront largely mirrored Study 1. Shifting from a hypothetical scenario in

Study 1, Study 2 allowed participants to respond to the racist post from condition materials (i.e., open-ended text responses), in addition to reporting on their anticipated social costs or benefits of confronting the same racist tweet they viewed others confront or not confront. We hypothesized that participants would expect fewer social costs and greater benefits in the confronting norm condition compared to the conditions where confronting was not the norm. Further, we assessed participants' general prejudice confrontation intentions as in Study 1, and, novel to Study 2, their intentions to confront the originally posted anti-Asian tweet and a novel anti-Asian tweet by a separate poster. We hypothesized greater prejudice confrontations across these three measures in the confrontation norm condition compared to both other conditions. However, broad confrontation intentions may vary compared to prejudice confrontation intentions of a specific instance of bias, as people often incorrectly forecast their prejudice confrontation likelihood when prejudicial encounters occur (Hurd et al., 2022; Kawakami et al., 2009; Swim & Hyers, 1999).

Study 2 included various manipulation checks to demonstrate that a social norm was successfully being manipulated as present or absent. When a behavior is considered normative, it is often perceived as more socially acceptable and appropriate to perform (e.g., Cialdini et al., 1990; Cialdini & Trost, 1998; Rimal et al., 2005). As such, we hypothesized that participants would perceive greater response appropriateness of individuals' responses to a racist tweet when there was a descriptive confrontation norm in place (i.e., Confronting Norm, 100%) compared to the absence of confrontation norms (No Confronting Norm, 0%; No Confronting Norm, 17%). Finally, as we were now examining confrontations of an anti-Asian comment, participants' own attitude towards Asian Americans was assessed and included as a covariate in the analyses.

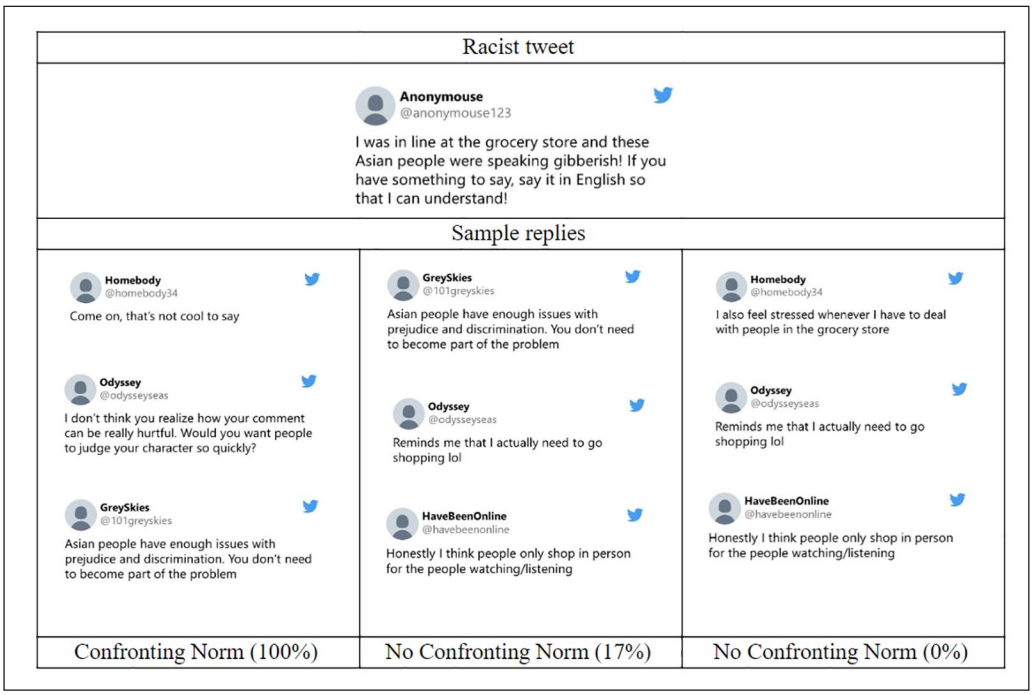
Method

Participants. Participants were recruited from Prolific in exchange for compensation. An a priori power analysis in G*Power (Faul et al., 2007) for a three-cell, between-subjects analysis of covariance (ANCOVA) to determine the desired sample size with 95% power to detect a medium effect ($d = 0.25$) revealed a desired sample size of 251. To allow for potential exclusions, 290 participants who identified as non-Hispanic White residing in the US were recruited from Prolific during May 2023 in exchange for compensation. Participants were excluded from analyses for not identifying as White ($n = 7$), non-Hispanic ($n = 3$), and failing at two or more attention checks ($n = 6$), leaving an analytic sample of 274. All participants identified as non-Hispanic White (see Table 1 for demographics).

Procedures. Participants were invited to participate in an online survey about perceptions of social interactions. Across all conditions, participants read an ostensible anti-Asian tweet (i.e., "I was in line at the grocery store and these Asian people were speaking gibberish! If you have something to say, say it in English so that I can understand!"; adapted from Meyers et al., 2020) before reading six replies from six different Twitter users. Tweets depicted raceless and genderless icons, with anonymized user handles (e.g., Twitter user Homebody; see Supplemental Material for full condition materials). Participants were randomly assigned to one of three conditions for the six replies (see Figure 3).

Participants in the Confronting Norm (100%) condition saw six Twitter users confront the racist tweet in either a neutral (e.g., "Come on, that's not cool to say") or educational style (e.g., "Asian people have enough issues with prejudice and discrimination. You don't need to become part of the problem"; Chaney & Sanchez, 2022). In the No Confronting Norm (17%) condition, participants only saw one out of six Twitter users confront the racist tweet. The single confronter used an educational style. The nonconfronters commented on nonracist aspects of the racist tweet (e.g., "Reminds me that I actually need to go shopping lol"). Participants in the No Confronting Norm (0%) condition saw none of the six Twitter users confront the racist tweet, and instead all commented on nonracist aspects of the tweet, as in the No

Figure 3. Descriptive norm condition materials: Study 2.



Note. Sample replies are provided (see Supplemental Material for full condition materials).

Confronting Norm (17%) condition. Note that in a pilot study with 185 non-Hispanic White undergraduates during April 2023, perceived racism of the original poster did not significantly differ across conditions and was viewed as highly racist on a 7-point scale (1 = *not at all*, 7 = *very*); $F(2, 181) = 0.32$, $p = .727$, $d = 0.13$ (Confronting Norm 100%: $M = 6.18$, $SE = 0.11$; No Confronting Norm 17%: $M = 6.26$, $SE = 0.11$; No Confronting Norm 0%: $M = 6.31$, $SE = 0.12$).

Next, participants were presented with the condition materials for the twitter replies and were asked to report on their perceived appropriateness before reporting on a measure indicating perceived descriptive confrontation norm. Participants were then asked an open-ended question about what they would reply in a tweet, if anything, to the perpetrator's racist tweet before completing a single item assessing likelihood of confronting the racist tweet from the condition materials they reviewed. Next, participants completed the following scales from Study 1: measures of anticipated social costs ($\alpha = .88$) and social benefits ($\alpha = .88$) of confronting prejudice,[2] adapted to consider confronting the racist tweet in the manipulation. Then participants completed an open-ended response to the new racist tweet, intentions to confront a new racist tweet (single item), and general confrontation intentions ($\alpha = .92$; for additional measures, see the Supplemental Material).[3] Participants reported on demographic questions as well as their warmth/coldness toward Asian people on a 0–100 slider scale (e.g., Duckitt & Sibley, 2007) before being debriefed and compensated.

Materials

Perceived response appropriateness. Participants completed three items (developed by the authors of the present study) assessing the extent to which the responses toward the first tweet were

appropriate (e.g., "To what extent were the previous tweets . . . responses that were socially acceptable"; $\alpha = .97$). Items were rated on a 7-point Likert scale (1 = *Not at all likely*, 7 = *Very likely*).

Perceived descriptive prejudice confrontation norm. Participants completed three items that assessed a descriptive norm of prejudice confrontation among those who replied to the racist tweet (e.g., "Many people confront prejudice or discrimination"; $\alpha = .93$; adapted from Park & Smith, 2007). Items were rated on a 7-point Likert scale (1= *Strongly agree*, 7 = *Strongly disagree*).

Dichotomous coding of confrontation responses to first and new racist tweet. Two trained research assistants, blind to condition and hypotheses, coded participants' open-ended responses for the presence of a confrontation (1) or the absence of confrontation (0). Discrepancies between the two research assistants were resolved with a third rater. Interrater agreement was calculated as percentage agreement, with 95% agreement reached for both the first and new racist tweet responses. Confrontations were defined as cases when the participant, in their open-ended responses, "does condemn, disapprove of, or 'call out' the person for their prejudice in a direct confrontation, 'That's pretty racist,' or an indirect confrontation, 'Not nice to say.'"

Intentions to confront the first racist tweet. Participants across conditions were presented with the image of the initial pilot-tested racist tweet loosely adapted from anti-Asian rhetoric in the US (e.g., Meyers et al., 2020). Then, participants were asked to rate on a 7-point Likert scale (1= *Not at all likely*, 7 = *Very likely*) how likely they were "to confront, or indicate disagreement..." with the individual who wrote the tweet.[4]

Intentions to confront a new racist tweet. Participants were shown a new, non-pilot-tested anti-Asian tweet from another user (i.e., "Asian people are so good at making cheap food . . . guess I'm taking applications for an Asian maid!"). Then, participants rated on a 7-point scale (1 = *Not at all likely*, 7 = *Very likely*) how likely they were "to confront, or indicate disagreement..." with the individual who wrote the tweet.

Warmth toward Asian people. Participants rated on a slider scale from 0 (*Very cold/Negative*) to 100 (*Very warm/Positive*) how warmly they felt toward Asian people (e.g., Duckitt & Sibley, 2007).

Results

Primary analyses were conducted as one-way ANCOVAs, controlling for participants' warmth toward Asian people. LSD post hoc tests were conducted. Participants' warmth toward Asian people did not differ by condition, $F(2, 271) = 1.22$, $p = .296$, $d = 0.19$ ($M = 80.71$, $SE = 1.24$). See Table 3 for ANCOVA results and descriptive statistics by condition.

Pearson's chi-squares of independence were conducted to examine the association of condition with the dichotomous coding of confrontation from participants' open-ended responses to both the first and the new racist tweet.

Manipulation checks. There was a significant effect of condition on perceived response appropriateness of the six Twitter users' replies (see Table 3). Participants perceived greater response appropriateness in the Confronting Norm (100%) condition than in the No Confronting Norm (17%) condition, $p < .001$, $d = 1.77$, 95% $CI_{meandiff}$[1.94, 2.76], and in the No Confronting Norm (0%) condition, $p < .001$, $d = 2.37$, 95% $CI_{meandiff}$[2.82, 3.65]. Further, participants reported greater response appropriateness in the No Confronting Norm (17%) than in the No Confronting Norm (0%) condition, $p < .001$, $d = 0.59$, 95% $CI_{meandiff}$[0.48, 1.29].

There was also a significant effect of condition on perceived descriptive prejudice confrontation norms among Twitter users. Participants perceived greater confrontation norms in the Confronting Norm (100%) condition than in the No Confronting Norm (17%) condition, $p < .001$, $d = 1.19$, 95% $CI_{meandiff}$[1.20, 2.00], and in the No Confronting Norm (0%) condition, $p <$

Table 3. ANCOVA results: Study 2.

	Condition effect			Confronting Norm (100%)	No Confronting Norm (17%)	No Confronting Norm (0%)
Outcome	$F(2, 270)$	p	d	M (SE)	M (SE)	M (SE)
Perceived appropriateness	125.65	< .001	1.93	6.22 (0.15)	3.87 (0.15)	2.66 (1.19)
Perceived descriptive norm	78.22	< .001	1.53	5.37 (0.15)	3.77 (0.14)	2.82 (0.14)
Social costs	1.69	.186	0.22	3.73 (0.10)	3.86 (0.10)	3.99 (0.10)
Social benefits	3.52	.031	0.32	3.86 (0.10)	3.53 (0.10)	3.58 (0.10)
First tweet confrontation intentions	0.26	.773	0.09	3.95 (0.22)	3.74 (0.22)	3.90 (0.22)
New tweet confrontation intentions	1.49	.227	0.21	3.50 (0.21)	2.99 (0.21)	3.18 (0.21)
General confrontation intentions	2.41	.092	0.27	4.36 (0.18)	3.80 (0.18)	3.98 (0.18)

Note. Perceived appropriateness, $F(2, 269)$. ANCOVA = analysis of covariance.

.001, $d = 1.97$, 95% $CI_{meandiff}$[2.14, 2.95]. Further, participants reported greater descriptive confrontation norms in the No Confronting Norm (17%) condition than in the No Confronting Norm (0%) condition, $p < .001$, $d = 0.65$, 95% $CI_{meandiff}$[0.55, 1.35].

Anticipated social costs and benefits of confronting. There was no significant main effect of condition on anticipated social costs of confronting the first racist tweet. Participants anticipated low to moderate social costs of confronting this tweet across conditions.

There was a significant main effect of condition on anticipated social benefits of confronting the first racist tweet. Participants anticipated greater social benefits in the Confronting Norm (100%) condition than in the No Confronting Norm (17%) condition, $p = .014$, $d = 0.36$, 95% $CI_{meandiff}$[0.07, 0.64], and in the No Confronting Norm (0%) condition, $p = .039$, $d = 0.49$, 95% $CI_{meandiff}$[0.16, 0.59]. Participants reported similar social benefits in the No Confronting Norm (17%) condition as in the No Confronting Norm (0%) condition, $p = .700$, $d = 0.08$, 95% $CI_{meandiff}$[−0.34, 0.23].

Dichotomous coding of confrontation responses to the first and new racist Tweet. Pearson's chi-square revealed no significant effect of condition on participants' coded confrontation toward the first racist tweet, $\chi^2(2, n = 274) = 0.19$, $p = .909$, Cramer's $v = .03$. In the Confronting Norm (100%) condition, 69.7% ($n = 62$) of participants confronted, compared to 67.0% ($n = 63$) in the No Confronting Norm (17%) condition, and 67.0% ($n = 61$) in the No Confronting Norm (0%) condition.

Similarly, there was no significant effect of condition on coded confrontations toward the new racist tweet, $\chi^2(2, n = 274) = 2.81$, $p = .246$, Cramer's $v = .10$. In the Confronting Norm (100%) condition, 52.8% ($n = 47$) of participants confronted, compared to 62.8% ($n = 59$) in the No Confronting Norm (17%) condition and 51.6% ($n = 47$) in the No Confronting Norm (0%) condition.

Confrontation intentions. There were no significant main effects of condition on intentions to confront the first racist tweet, the new racist tweet, or general intentions to confront prejudice in the future. Participants had somewhat moderate intentions to confront prejudice across measures and conditions.

Mediations. Parallel mediation analyses were conducted as in Study 1 while controlling for participants' attitude toward Asian Americans. Two

analyses were conducted, the first with general confrontation intentions, and the second with a composite measure of anti-Asian prejudice confrontation intentions as the outcome (from the first and the new racist tweet confrontation intentions).[4] See Figure 5. Indirect effects statistics are outlined in Table 4.

The indirect effects of Contrast 1 (100% vs. 17% confronting) and Contrast 2 (100% vs. 0% confronting) on anti-Asian prejudice confrontation intentions and general confrontation intentions were not significant via perceived social costs, but were significant through perceived social benefits. Participants in the Confronting Norm (100%) condition perceived significantly greater benefits of confronting prejudice relative to participants in the No Confronting Norm (17%) condition and in the No Confronting Norm (0%) condition, which was in turn associated with greater anti-Asian and general prejudice confrontation intentions.

Discussion

Study 2 successfully manipulated prejudice confrontation norms via group behavior. When everyone confronted, participants perceived a greater descriptive norm of prejudice confrontations, and perceived prejudice confrontations as more appropriate compared to when no norm was present. Contrary to predictions, participants also reported greater response appropriateness in the No Confronting Norm (17%) condition than in the No Confronting Norm (0%) condition. This may be due, in part, to perceived lower social acceptability of publicly expressing extreme and overt displays of racism in the US (e.g., Crandall et al., 2002). Thus, participants perceived the collection of social media responses to be more socially acceptable when a single poster condemned overt anti-Asian racism, compared to when no social media posters addressed the overt anti-Asian racism. Indeed, one might interpret the No Confronting Norm (0%) condition as the posters, who all failed to confront the perpetrator, signaling agreement with the perpetrator's racism (e.g., Zapata et al., 2024).

As in Study 1, participants forecasted greater social benefits of confronting prejudice when a confronting norm was in place compared to absent, though in response to a specific racist incident in Study 2. Further replicating Study 1, perceived social benefits significantly mediated general intentions to confront and, novel to Study 2, intentions to confront anti-Asian prejudice. However, no significant effect of prejudice confrontation norms on confrontation intentions or coded open-ended responses emerged. Further, contrary to Study 1, Study 2 found no differences in anticipated social costs of confronting prejudice in the presence or absence of a prejudice confrontation norm.

The use of anonymous online users was meant to isolate the effects of a prejudice confrontation social norm message from the effects of perpetrator's and potential confronters' identities. Yet, perceived social costs and benefits are likely to differ when racism and confrontations are enacted by anonymous people compared to known individuals. Further, social norm manipulations may be more effective at shifting behavior when the behavior is the ingroup social norm (Paluck & Shepherd, 2012; Rimal et al., 2005). Thus, Study 3 incorporated online users who matched participants' racial ingroup members (e.g., White people).

Study 3

Study 3 sought to replicate and extend upon findings with perceived online descriptive confrontation norms in Study 2 by incorporating visual stimuli of White men and women in the profiles of the social media posts. This addition sought to add realism to the manipulation of social media interactions, where users are likely to convey their race and gender via profile pictures (e.g., Hurd et al., 2022; Meyers et al., 2020). Further, incorporating visual stimuli of social media users tailors the descriptive prejudice confrontation norms to explicitly reference White participants' racial ingroup members upholding the depicted norm. In turn, the explicit reference of one's racial ingroup may impact confrontation intentions

Table 4. Mediation analyses indirect effects: Study 2.

Contrast	Outcome variable	Mediator	*B*	*SE*	95% CI_{boot} LL	95% CI_{boot} UL
Contrast 1: 0 = Confronting Norm (100%) vs. 1 = No Confronting Norm (17%)	Anti-Asian prejudice confrontations	Social costs	0.05	0.05	−0.05	0.16
		Social benefits	−0.25	0.11	−0.48	−0.04
	General confrontation intentions	Social costs	0.04	0.04	−0.04	0.12
		Social benefits	−0.28	0.13	−0.54	−0.05
Contrast 2: 0 = Confronting Norm (100%) vs. 1 = No Confronting Norm (0%)	Anti-Asian prejudice confrontations	Social costs	0.09	0.06	−0.01	0.22
		Social benefits	−0.21	0.10	−0.41	−0.04
	General confrontation intentions	Social costs	0.07	0.05	−0.01	0.18
		Social benefits	−0.24	0.12	−0.48	−0.02

through expectations of social costs and benefits of confronting formed from a normalized behavior exhibited from one's ingroup. Study 3 was preregistered (https://osf.io/6g7yp/?=61be4d8fac824dcabb576a3935cc8d78).

Our primary hypotheses mirrored Study 2 and included perceived confrontation appropriateness as a manipulation check, as well as anticipated social costs and benefits, and intentions to confront the first and new racist tweets.[5] Participants' general intentions to confront prejudice were preregistered as a secondary outcome.

Method

Participants. An a priori power analysis in G*Power (Faul et al., 2007) for a three-cell between-subjects ANCOVA revealed a desired sample size of 432 to detect a small effect (d = 0.30) with 80% power. To allow for potential exclusions, 471 participants who identified as non-Hispanic White residing in the US were recruited from Prolific during July 2023. Eighteen participants were excluded from analyses for not identifying as non-Hispanic White and 13 were excluded for failing two or more attention checks, leaving an analytic sample of 450. See Table 1 for demographics.

Procedures. Procedures, materials, and study design were identical to Study 2, except tweets now depicted White men and women with neutral expressions, using randomly selected stimuli from the Chicago Face Database (CFD; Ma et al., 2015). Across all conditions (i.e., Confronting Norm [100%]; No Confronting Norm [17%]; No Confronting Norm [0%]), participants read the same ostensible anti-Asian tweet as in Study 2, now from a White man, before reading the same six replies from six different Twitter users as in Study 2. However, these responses now came from three White women and three White men Twitter users with user handles invoking prototypical White names (e.g., Ryan Allen, @ryallen567; see Supplemental Material for full condition materials). In the No Confronting Norm (17%), the sole confronter was a White woman.[6]

As in Study 2, participants completed measures of perceived response appropriateness (α = .96), perceived descriptive prejudice confrontation norm (α = .91), an open-ended response to the first racist tweet, confrontation intentions toward the first racist tweet, measures of anticipated social costs (α = .90) and social benefits of confronting the first anti-Asian racist tweet (α = .85), open-ended response to the new racist tweet,[7] confrontation intentions toward the new anti-Asian racist tweet, and general confrontation intentions (α = .90). Participants then reported their attitude towards Asian Americans and demographics before being debriefed and compensated. Interrater agreement was calculated as

percentage agreement for the coded open-ended responses; 88% agreement was reached by research assistants for both the first and new racist tweet responses (more than 80% agreement indicates reasonable reliability; e.g., Miles et al., 2014; Oswald et al., 2022). For additional findings on measures not reported here, see Supplemental Material.[8]

Results

Analyses were conducted as one-way ANCOVAs, again controlling for participants' warmth toward Asian people. LSD post hoc tests were conducted (see Table 5). Participants' warmth toward Asian people did not significantly differ by condition, $F(2, 447) = 0.08$, $p = .924$, $d < 0.01$ ($M = 80.87$, $SE = 0.86$). Pearson's chi-squares of independence were calculated to examine the association of condition with the dichotomous coding of confrontation from participants' open-ended responses to both the first and new racist tweets.

Manipulation checks. Mirroring Study 2, there was a significant effect of condition on perceived response appropriateness of the replies and perceived descriptive norm. Participants perceived greater response appropriateness in the Confronting Norm (100%) condition than in the No Confronting Norm (17%) condition, $p < .001$, $d = 1.52$, 95% $CI_{meandiff}$[1.75, 2.39], and in the No Confronting Norm (0%) condition, $p < .001$, $d = 2.05$, 95% $CI_{meandiff}$[2.62, 3.26]. Further, participants reported greater response appropriateness in the No Confronting Norm (17%) than in the No Confronting Norm (0%) condition, $p < .001$, $d = 0.62$, 95% $CI_{meandiff}$[0.56, 1.19].

Participants perceived a greater descriptive prejudice confrontation norm in the Confronting Norm (100%) condition than in the No Confronting Norm (17%) condition, $p < .001$, $d = 0.93$, 95% $CI_{meandiff}$[0.96, 1.61], and in the No Confronting Norm (0%) condition, $p < .001$, $d = 1.57$, 95% $CI_{meandiff}$[1.90, 2.55]. Further, participants reported a greater descriptive prejudice confrontation norm in the No Confronting Norm (17%) condition than in the No Confronting Norm (0%) condition, $p < .001$, $d = 0.63$, 95% $CI_{meandiff}$[0.61, 1.26].

Anticipated social costs and benefits. There was a significant main effect of condition on anticipated social costs of confronting the first racist tweet (see Figure 4). Participants anticipated fewer social costs of confronting in the Confronting Norm (100%) condition than in the No Confronting Norm (17%) condition, $p = .017$, $d = 0.26$, 95% $CI_{meandiff}$[−0.52, −0.05], and in the No Confronting Norm (0%) condition, $p < .001$, $d = 0.52$, 95% $CI_{meandiff}$[−0.76, −0.29]. Further, participants reported fewer social costs in the No Confronting Norm (17%) condition than in the No Confronting Norm (0%) condition, $p = .046$, $d = 0.24$, 95% $CI_{meandiff}$[−0.48, −0.01].

There was a significant main effect of condition on anticipated social benefits of confronting the first racist tweet. Participants anticipated greater social benefits in the Confronting Norm (100%) condition than in the No Confronting Norm (0%) condition, $p = .012$, $d = 0.29$, 95% $CI_{meandiff}$[0.05, 0.44]. Further, participants reported greater social benefits in the No Confronting Norm (17%) condition than in the No Confronting Norm (0%) condition, $p = .004$, $d = 0.35$, 95% $CI_{meandiff}$[0.09, 0.47]. However, there was no significant difference in social benefit expectations of confronting between the Confronting Norm (100%) condition and the No Confronting Norm (17%) condition, $p = .704$, $d = 0.05$, 95% $CI_{meandiff}$[−0.23, 0.15].

Dichotomous coding of confrontation responses to first and new racist tweets. A Pearson's chi-square revealed a significant effect of condition on participants' coded confrontation toward the first racist tweet, $\chi^2(2, n = 450) = 7.95$, $p = .019$, Cramer's $v = .13$. Bonferroni corrected post hoc z-tests revealed that participants were most likely to confront in the No Confronting Norm (0%) condition (80.4%, $n = 119$) than in the Confronting Norm (100%) condition (66.7%, $n = 100$) and in the No Confronting Norm (17%) condition (69.1%, $n = 105$), which did not significantly differ from each other.

Table 5. ANCOVA results: Study 3.

	Condition effect			Confronting Norm (100%)	No Confronting Norm (17%)	No Confronting Norm (0%)
Outcome	$F(2, 446)$	p	d	M (SE)	M (SE)	M (SE)
Perceived appropriateness	171.21	< .001	1.76	6.01 (0.12)	3.95 (0.11)	3.07 (0.12)
Perceived descriptive norm	90.52	< .001	1.28	5.25 (0.12)	3.96 (0.12)	3.02 (0.12)
Social costs	9.58	< .001	0.41	3.67 (0.09)	3.95 (0.08)	4.19 (0.09)
Social benefits	4.96	.007	0.30	3.82 (0.07)	3.86 (0.07)	3.57 (0.07)
General confrontation intentions	0.99	.372	0.13	4.06 (0.13)	4.17 (0.13)	4.33 (0.13)
Confrontation intentions first tweet	1.39	.250	0.16	3.94 (0.17)	3.97 (0.17)	4.31 (0.17)
Confrontation intententions new tweet	2.06	.129	0.19	3.27 (0.17)	3.17 (0.17)	3.63 (0.17)

Note. $F(2,446)$ for outcomes except for perceived descriptive prejudice confrontation norm, $F(2, 445)$ and perceived response appropriateness, $F(2, 441)$, due to missing data. ANCOVA = analysis of covariance.

However, there was no significant effect of condition on coded confrontations toward the new racist tweet, $\chi^2(2, n = 450) = 4.99, p = .083$, Cramer's $v = .08$. Participants were equally likely to confront in the Confronting Norm (100%) condition (50.7%, $n = 76$), the No Confronting Norm (17%) condition (50.0%, $n = 76$), and the No Confronting Norm (0%) condition (61.5%, $n = 91$).

Confrontation intentions. There was no significant main effect of condition on participants' intentions to confront the first racist tweet, the new racist tweet, or general intentions to confront prejudice in the future. Participants had moderate intentions to confront prejudice across measures and conditions.

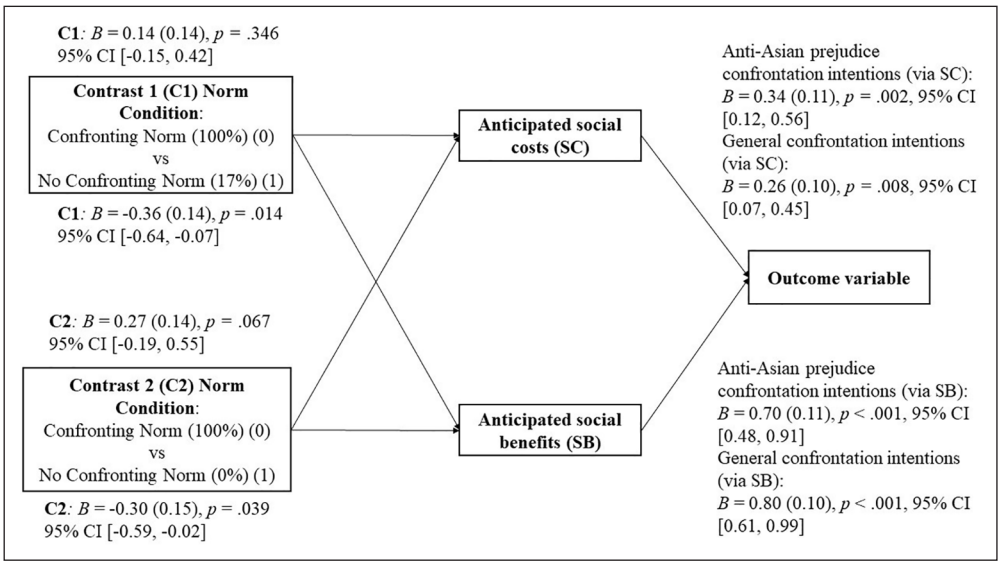
Mediations. Exploratory mediation analyses mirrored Study 2[9] (for path effects and statistics, see Figure 6). Indirect effects statistics are outlined in Table 6. The indirect effect of Contrast 1 (100% vs. 17% confrontation) on both anti-Asian prejudice confrontation intentions and general confrontation intentions was significant via perceived social costs, but not via perceived social benefits. Participants in the Confronting Norm (100%) condition perceived significantly fewer costs of confronting prejudice, which was in turn associated with greater prejudice confrontation intentions relative to participants in the No Confronting Norm (17%) condition.

The indirect effects of Contrast 2 (100% vs. 0% confronting) on both anti-Asian prejudice confrontation intentions and general confrontation intentions were significant via both perceived social costs and social benefits. Participants in the Confronting Norm (100%) condition perceived significantly greater social benefits and fewer social costs relative to participants in the No Confronting Norm (0%) condition, which was in turn related to greater intentions to confront prejudice.

Discussion

Study 3 expanded upon Study 2 by using raced and gendered perpetrators in the manipulation of perceptions of descriptive confrontation norms via observation of group behavior. Replicating findings from Study 2, participants perceived greater response appropriateness when at least one person confronted (17% and 100%), compared to when an anti-Asian tweet went unconfronted. When participants anticipated their own social costs and benefits of confronting the first racist tweet, both confronting conditions (100%, 17%) elicited fewer anticipated social costs and greater social benefits, compared to the absence

Figure 4. Mediation: Study 2.



Note. Unstandardized regression coefficients are reported. Standard errors are reported in parentheses.

of a norm. Though different from hypotheses, these findings suggest that ingroup members confronting prejudice, be it a wide ingroup norm upheld by multiple members or a single ingroup member confronting, mitigates perceived risks and boosts perceived benefits of confronting prejudice. Contrary to hypotheses, but consistent with Studies 1–2, online descriptive confrontation norms did not directly impact intentions to confront prejudice.

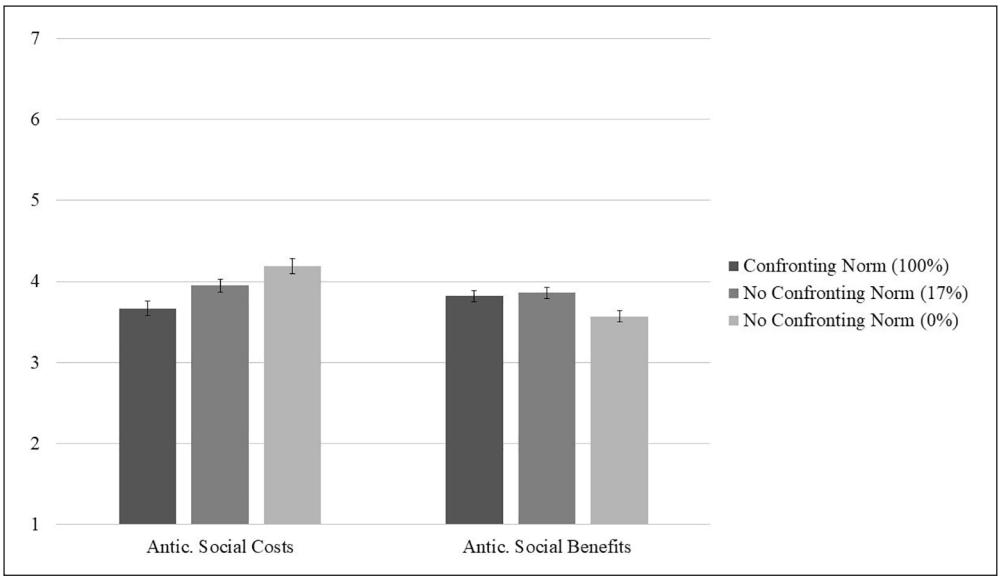
Indeed, there was only a significant effect of prejudice confrontation norms on coded open-ended responses to the first racist tweet. In the open-ended text, participants were more likely to confront in the No Confronting Norm (0%) condition than in the other descriptive confrontation norm conditions. This may suggest that participants were more inclined to confront when viewing responses indicating a "high prejudice" norm. Nevertheless, the indirect effect of a confronting norm on confrontation intentions was significant: fewer social costs and greater social benefits stemming from a descriptive norm of prejudice confrontation, relative to no one confronting prejudice, was associated with greater intentions to confront both the original and new anti-Asian tweets.

General Discussion

Despite prejudice confrontations seemingly being an effective tool for curbing prejudice in perpetrators and promoting belonging for marginalized groups (Chaney & Sanchez, 2018; Chu & Ashburn-Nardo, 2022; Czopp et al., 2006; Hildebrand et al., 2020; Li et al., 2024; Munger, 2017), prejudice confrontations remain a non-typical response to prejudice. Frequently identified barriers to confronting prejudice include concerns about social costs and limited anticipated social benefits. The present research sought to examine the utility of manipulating social norms of prejudice confrontations to shift these barriers, with the aim of promoting greater prejudice confrontation rates among White Americans.

When descriptive social norm information was communicated via group consensus information or ingroup behavior, White people expected fewer social costs of confronting prejudice generally (Study 1) and of confronting anti-Asian

Figure 5. Anticipated social costs and benefits by condition: Study 3.



Note. Error bars denote standard error.

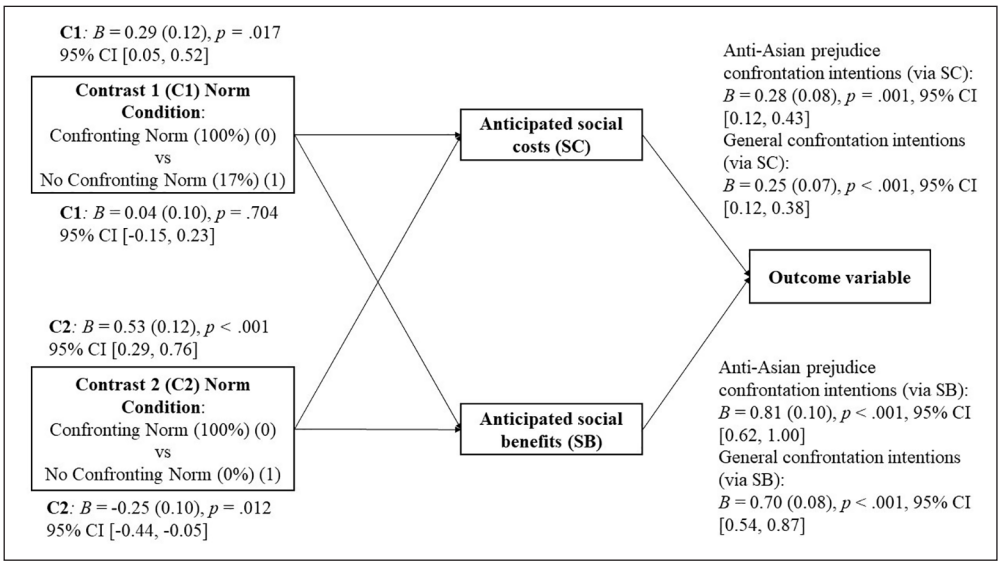
bias specifically (Study 3), compared to when no social norm of confronting prejudice was present. Notably, group behavior of anonymous online users did not significantly impact anticipated social costs of confronting prejudice (Study 2), perhaps suggesting the importance of felt similarity to those creating the social norm. Yet, regardless of how the descriptive norm information was communicated, anticipated social benefits were significantly greater when confronting was a norm, compared to when it was not (Studies 1–3). This may suggest that prejudice confrontation norms facilitate greater anticipated changes in perpetrators' future bias or in marginalized group members' anticipated safety, regardless of who is confronting prejudice. Future research should directly compare the utility of these varied means of social norm manipulation to discern such differences.

While the current studies document how descriptive confrontation norms (compared to the absence of such norms) may shift anticipated barriers to confronting, intention to confront prejudice was never directly impacted, only indirectly impacted (Studies 1–3). This may suggest that a one-time declaration of prejudice confrontation social norms needs to be supplemented with continual environmental features that be congruent with the norm, such as egalitarian policies and consequences for those who deviate from the norm (e.g., Álvarez-Benjumea, 2025; Douglas et al., 2024; Nelson et al., 2011), so that decisions to engage in prejudice confrontations align not only with the social norms of one's peers, but also the environment. This work did not examine social norms within an academic or organizational context, as previous social norm messaging work has investigated (e.g., Hurd et al., 2022; Li et al., 2024; Murrar et al., 2020). That is, it may be critical to adapt such an intervention to explicitly target not only the person, but also the context.

Confronter(s)' Identity and Ingroup Norms

Interestingly, White people who observed White social media users adhering to a complete descriptive confronting norm (100% social media users

Figure 6. Mediation: Study 3.



Note. Unstandardized regression coefficients are reported. Standard errors are shown in parentheses.

confront) or a single individual confronting prejudice (17% social media users confront) reported fewer social costs and greater social benefits of confronting anti-Asian bias (Study 3), compared to when a prejudiced comment was not confronted. These findings suggest that the more similar observed social referents adhering to confrontation social norms are to one's racial ingroup, the greater the benefits inferred from a prejudice confrontation, even when such norms are not upheld by all social referents (Rimal et al., 2005; Tankard & Paluck, 2016). Some research suggests that invoking a common, superordinate identity among religious ingroups and outgroups may also elicit greater support for messages that confront sectarian speech (Siegal & Badaan, 2020), suggesting another avenue of social referents for norm investigation. We encourage future research to directly compare the effect of social norm referent (ingroup, outgroup, no group knowledge, superordinate identity) with other social groups to determine if our findings are an ingroup phenomenon or one based on White Americans' privileged social status.

To date, previous work has found that Asian Americans find racist online posts more offensive when confronted by racial outgroup members (i.e., White confronters) than by racial ingroup members (i.e., Asian confronters, Studies 2–3; Meyers et al., 2020), but it has not examined how normative racial ingroup and outgroup behavior influences expectations of social costs or benefits of confronting. Hence, it is less clear if tailoring the descriptive confrontation norm to only those in one's ingroup would facilitate expectations of greater benefits and fewer costs. For example, descriptive social norms of only Asian individuals confronting anti-Asian bias might signal socially advantaged groups are not upholding the norm (e.g., White people not seen confronting, thus it would be expected that White people do not confront prejudice) and may not be effective or genuine allies. However, such representation of one's own ingroup confronting prejudice may signal that the environment is safe enough for such actions to occur with minimal costs (e.g., confrontation as empowering and eliciting autonomy; Chaney et al., 2015; Sanchez et al., 2016).

Table 6. Mediation analyses indirect effects: Study 3.

Contrast	Outcome variable	Mediator	*B*	*SE*	95% CI_{boot} LL	95% CI_{boot} UL
Contrast 1: 0 = Confronting Norm (100%) vs. 1 = No Confronting Norm (17%)	Anti-Asian prejudice confrontations	Social costs	0.08	0.04	0.01	0.17
		Social benefits	0.03	0.08	−0.13	0.18
	General confrontation intentions	Social costs	0.07	0.04	0.01	0.15
		Social benefits	0.03	0.07	−0.11	0.17
Contrast 2: 0 = Confronting Norm (100%) vs. 1 = No Confronting Norm (0%)	Anti-Asian prejudice confrontations	Social costs	0.15	0.05	0.05	0.26
		Social benefits	−0.20	0.08	−0.37	−0.05
	General confrontation intentions	Social costs	0.13	0.05	0.05	0.23
		Social benefits	−0.17	0.07	−0.32	−0.04

Confronting Anti-Asian Bias

The current work is one of few to examine prejudice confrontation of anti-Asian racism, as confrontation research predominantly examines anti-Black racism or sexism. We examined overt forms of anti-Asian bias, though we encourage future research to examine if descriptive confrontation norms of more subtle forms of anti-Asian bias may similarly impact anticipated social costs and benefits of confronting prejudice (e.g., Meyers et al., 2020; Ratcliff et al., 2023). This work examined anti-Asian bias in an effort to extend prejudice confrontation research to a less examined marginalized racial group. However, descriptive confrontation norms for one type of bias (here, anti-Asian) may similarly impact anticipated costs and benefits of confronting other types of bias due to lay beliefs about the generalized nature of prejudice (Chaney et al., 2016; Sanchez et al., 2017). Further, despite the prevalence of anti-Asian racism in the US (M. H. J. Lee et al., 2024; R. T. Lee et al., 2019), such bias has been generally recognized as undesirable behavior (Bašić et al., 2020; Crandall et al., 2002; Sommers & Norton, 2006). Descriptive norms of prejudice confrontation may reduce barriers to confronting biases that are generally deemed undesirable, rather than forms of bias that have not reached the same public recognition of being socially unacceptable (e.g., weight stigma, ageism; Crandall et al., 2002; Puhl & Heuer, 2009).

Future Directions

Though social norms of prejudice confrontation did not directly impact White people's intentions to confront prejudice generally (Study 1), or anti-Asian bias specifically (Study 2-3), such mechanisms did indirectly elicit greater intentions to confront prejudice (Studies 1-3), consistent with work suggesting high benefits and low costs facilitate the decision to confront prejudice (e.g., CPR model; Ashburn-Nardo & Karim, 2019; Ashburn-Nardo et al., 2008). Critically, the present research utilized online samples and assessed confrontation intentions within virtual spaces, which may not manifest as actual confrontation behavior in either physical or virtual spaces (e.g., Dickter &

Newton, 2013; Hurd et al., 2022). Thus, future research is needed to determine if social norm messaging may shift actual prejudice confrontation behaviors, moving beyond hypothetical or online scenarios. Social norm pressure may be greater in person compared to online, as social costs of not confronting prejudice may emerge in contexts where prejudice confrontation norms are high.

Only one study (Study 1) examined the application of injunctive norms of prejudice confrontation via a bogus scientific research article. Future work should further examine the consequences of perceived injunctive confrontation norms, for example, in social media spaces. Some social media sites explicitly adopt an injunctive norm around not posting "hate speech" in their community guidelines (e.g., Matias, 2019), though this may be inconsistently enacted across time and platforms. Further, reviews of how social norms can facilitate behavior change suggest that broad societal injunctive norms that require refraining from biased behavior and engaging in more pro-social behavior may only influence behavior change through consistent expectations of social backlash or legal systems of punishment for violators of the norm (e.g., Álvarez-Benjumea, 2025; Douglas et al., 2024). For example, in a field study with a forum page, first-time commenters that were exposed to injunctive egalitarian norms of community behavior (e.g., informed of the unacceptable behaviors, consequences of breaking the rules, reports that everyone agrees with the rules), compared to those not exposed to norm messaging, were more likely to engage in egalitarian posting behaviors (Matias, 2019). Further, high-status social referents that address norm violators facilitate effective bias reduction online, including up to 1 month later (e.g., Munger, 2017; Siegal & Badaan, 2020). This may suggest that a combination of descriptive and injunctive confrontation norms, rather than a single norm, may be the most effective in promoting prejudice confrontation (e.g., Mauduy et al., 2022). Therefore, when integrating our findings from online descriptive confrontation norms, we contend that injunctive confronting norms may be most effective when (a) the social referent is observed personally confronting prejudice (e.g., Studies 2–3) and (b) the confronted person faces some social backlash for being biased (e.g., shamed by a high-status user with social influence or banned from the social media site; Matias, 2019; Munger, 2017; Siegel & Badaan, 2020).

Though we have documented how descriptive confrontation norms may mitigate some barriers, additional ones, such as not knowing how to confront (e.g., Ashburn-Nardo & Karim, 2019; Ashburn-Nardo et al., 2008), might only be addressed through practice or repeated observation of how individuals confront a wide variety of prejudices. Further, interpersonal displays of prejudice will only be confronted if they are first recognized as being prejudicial (Ashburn-Nardo & Karim, 2019; Ashburn-Nardo et al., 2008, 2014). Indeed, people tend to confront more overt forms of prejudice (Dickter, 2012; Dickter et al., 2012; Meyers et al., 2020), perhaps in part because hostile and aggressive stereotypical beliefs about marginalized groups are generally deemed socially inacceptable in public spaces (i.e., justification–suppression model of prejudice; Crandall et al., 2002, Crandall & Eshleman, 2003). Lastly, past research has highlighted the importance of an affirmed prejudice confrontation (i.e., one or more people confronting prejudice after an initial confrontation) in signaling a context's egalitarian norm and promoting identity safety for marginalized groups (Hildebrand et al., 2020; Li et al., 2024). Social norm messaging about prejudice confrontations may thus be useful in signaling an egalitarian climate and promoting identity safety for marginalized groups if it can promote greater prejudice confrontation rates, and thus more frequent occurrences of affirmed confrontations. It is imperative to assess how marginalized group members perceive such bias reduction efforts, as confrontations may be perceived as disingenuous (e.g., Burns & Granz, 2023; Thai & Nylund, 2024).

Conclusion

The present study expanded prejudice confrontation research by incorporating social norm

messaging as a vehicle to reduce common barriers for White people to confront prejudice. Descriptive social norms indirectly elicited confrontation intentions through reduced expectations of social costs (Study 3) and increased social benefits of confronting prejudice (Studies 1–3), relative to the absence of a norm. The current research demonstrates that descriptive norms of prejudice confrontation—both as group consensus information (Study 1) and observed interactions amongst White social media users (Study 3)—reduce social costs and increase social benefits of confronting prejudice generally (Study 1) and anti-Asian racism specifically (Studies 2–3).

Authors' Note

The authors have moved to a new institution since completing this research, this being University at Buffalo, SUNY. The institution where the research was conducted was University of Connecticut, Department of Psychological Sciences.

Consent to Participate and Consent for Publication

All participants indicated informed consent before taking part in our studies, including consent for their data to be de-identified and shared in conferences and publications.

Data Availability

All data and materials are provided on the OSF (https://osf.io/v7kwz/?=4f6412282a644105b3c808d4b6c563ed) or in the Supplemental Material.

Ethical Considerations

All participants in this research were treated in accordance with APA guidelines on ethical treatment of human subjects, and the studies were approved by the University of Connecticut's Institutional Review Board.

Funding

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

ORCID iDs

Izilda Pereira-Jorge https://orcid.org/0000-0002-4798-8779
Kimberly E. Chaney https://orcid.org/0000-0001-6450-9488

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Additional measures in the Supplemental Material include potential covariates or moderators, including perceived value of confrontation, whom individuals would confront (e.g., a friend, an acquaintance), estimates of harm toward marginalized groups, metaracism beliefs, and social dominance orientation (SDO).
2. Items were similar to Study 1, with 16 (out of the original 17 items) for social costs, and 14 (out of 17) for social benefits, to best match costs and benefits of online social interactions.
3. Additional supplemental measures mirror previous supplemental measures in the pilot study, including additional manipulation checks such as perceived percentage of confronters. Measures also include perceived confronter social costs and social benefits, value of confrontations, perceived advantaged group's and marginalized group's confrontation frequency (composite), lay beliefs of prejudice origins, SDO, participants' warmth toward confronters, and metaracism beliefs.
4. Parallel mediations were also conducted for the coded confrontation responses to the first and new racist tweets. Mirroring the results of confrontation intentions and anti-Asian confrontation intentions, there was a significant indirect effect of Contrasts 1 and 2 via anticipated social benefits, but not social costs. Participants in the Confronting Norm (100%) condition perceived significantly greater benefits of confronting prejudice relative to participants in the No Confronting Norm (17%) condition and in the No Confronting Norm (0%) condition, which in turn was associated with greater likelihood of confronting the perpetrator of both the first and new racist tweets. See Supplemental Material for full results.
5. We also preregistered primary hypotheses for perceptions of social costs and benefits for social

media users who confronted the original racist post in the manipulation conditions (i.e., users in the Confronting Norm [100%] and in the No Confronting Norm [17%] condition). We asked participants to indicate how likely the target confronter (from either the 100% or 17% condition) was to incur social costs or benefits for confronting the first racist post. We hypothesized that participants would anticipate a confronter would face fewer social costs and greater social benefits in the Confronting Norm (100%) than in the No Confronting Norm (17%) condition. Because perceived social costs and benefits for confronters were highly correlated to participants' own anticipated social costs and benefits of confronting ($rs > .63$, $ps < .001$), this measure was removed from the main manuscript and placed in the Supplemental Material with other supplemental secondary hypotheses.

6. This is a deviation from the preregistration where we indicated the confronter would be a man.
7. Generally, pilot test results from April 2025 indicate the participants reported the new racist tweet as being racist ($M = 4.98$, $SE = 0.27$) on a 7-point Likert scale (1 = *Not at all racist*, 7 = *Very racist*), and perceived the perpetrator as somewhat racist ($M = 4.26$, $SE = 0.23$) on a 7-point Likert scale (1 = *Not at all true*, 7 = *Very true*). When asked about the difficulty of determining the offensiveness of the post, participants indicated low difficulty ($M = 2.53$, $SE = 0.25$) on a 7-point Likert scale (1 = *Not at all difficult*, 7 = *Very difficult*).
8. Additional measures in the Supplemental Material include perceived confronter social costs and benefits, value of confrontations, perceived advantaged group's and marginalized group's confrontation frequency (composite), perpetrator's perceived prejudice intentionality, lay beliefs of prejudice origins, SDO, etc.
9. Parallel mediations were also conducted for the coded confrontation responses to the first and new racist tweets. There was a significant indirect effect of Contrast 2 via anticipated social benefits, but not social costs. No significant indirect effects were found via social costs with Contrast 1 or Contrast 2. Participants in the Confronting Norm (100%) condition perceived significantly greater benefits of confronting prejudice relative to participants in the No Confronting Norm (0%) condition, which was in turn associated with greater likelihood of confronting the first and new racist tweet in their open-ended responses. See Supplemental Material for full results.

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 & Intergroup Relations*, *22*(5), 655–672. https://doi.org/10.1177/1368430218756493

Álvarez-Benjumea, A. (2025). Social norms and the expression of prejudice: How the norm changes. *Current Opinion in Psychology*, *62*, Article 101974. https://doi.org/10.1016/j.copsyc.2024.101974

Ashburn-Nardo, L., Blanchar, J. C., Petersson, J., Morris, K. A., & Goodwin, S. A. (2014). Do you say something when it's your boss? The role of perpetrator power in prejudice confrontation. *Journal of Social Issues*, *70*(4), 615–636. https://doi.org/10.1111/josi.12082

Ashburn-Nardo, L., & Karim, M. F. A. (2019). The CPR model: Decisions involved in confronting prejudiced responses. In R. K. Mallett & M. J. Monteith (Eds.), *Confronting prejudice and discrimination* (pp. 29–47). Academic Press.

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

Bašic, Z., Falk, A., & Kosse, F. (2020). The development of egalitarian norm enforcement in childhood and adolescence. *Journal of Economic Behavior & Organization*, *179*, 667–680. https://doi.org/10.1016/j.jebo.2019.03.014

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

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 & Intergroup Relations*, *26*(2), 453–475. https://doi.org/10.1177/13684302211059699

Chaney, K. E., & Chasteen, A. L. (2023). Do beliefs that older adults are inflexible serve as a barrier to racial equality? *Personality and Social Psychology Bulletin*, *50*(8), 1151–1166. https://doi.org/10.1177/01461672231159767

Chaney, K. E., O'Dea, C. J., & Pereira-Jorge, I. (2025). From confronted to confronter? Examining the enduring effects of prejudice confrontations. *Group Processes & Intergroup Relations*, *28*(4), 908–930. https://doi.org/10.1177/13684302241309872

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

Chaney, K. E., & Sanchez, D. T. (2022). Prejudice confrontation styles: A validated and reliable measure of how people confront prejudice. *Group Processes & Intergroup Relations*, *25*(5), 1333–1352. https://doi.org/10.1177/13684302211005841

Chaney, K. E., Sanchez, D. T., & Remedios, J. D. (2016). Organizational identity safety cue transfers. *Personality and Social Psychology Bulletin*, *42*(11), 1564–1576. https://doi.org/10.1177/0146167216665096

Chaney, K. E., Young, D. M., & Sanchez, D. T. (2015). Confrontation's health outcomes and promotion of egalitarianism (C-HOPE) framework. *Translational Issues in Psychological Science*, *1*(4), 363–371. https://doi.org/10.1037/tps0000042

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/10.1016/j.jesp.2022.104337

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence*, *1*(1), 3–15. https://doi.org/10.1080/15534510500181459

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. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological bulletin*, *129*(3), 414–446. https://doi.org/10.1037/0033-2909.129.3.414

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

Czopp, A. M. (2019). The consequences of confronting prejudice. In R. K. Mallett & M. J. Monteith (Eds.), *Confronting prejudice and discrimination* (pp. 201–221). Academic Press.

Czopp, A. M., & Monteith, M. J. (2003). Confronting prejudice (literally): Reactions to confrontations of racial and gender bias. *Personality and Social Psychology Bulletin*, *29*(4), 532–544. https://doi.org/10.1177/0146167202250923

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

De Souza, L., & Schmader, T. (2022). The misjudgment of men: Does pluralistic ignorance inhibit allyship? *Journal of Personality and Social Psychology*, *122*(2), 265–285. https://doi.org/10.1037/pspi0000362

Dickter, C. L. (2012). Confronting hate: Heterosexuals' responses to anti-gay comments. *Journal of Homosexuality*, *59*(8), 1113–1130. https://doi.org/10.1080/00918369.2012.712817

Dickter, C. L., Kittel, J. A., & Gyurovski, I. I. (2012). Perceptions of non-target confronters in response to racist and heterosexist remarks. *European Journal of Social Psychology*, *42*(1), 112–119. https://doi.org/10.1002/ejsp.855

Dickter, C. L., & Newton, V. A. (2013). To confront or not to confront: Non-targets' evaluations of and responses to racist comments. *Journal of Applied Social Psychology*, *43*(S2), E262–E275. https://doi.org/10.1111/jasp.12022

Douglas, B. D., Holley, K., Isenberg, N., Kennedy, K. R., & Brauer, M. (2024). Social sanctions in response to injunctive norm violations. *Current Opinion in Psychology*, *59*, Article 101850. https://doi.org/10.1016/j.copsyc.2024.101850

Duckitt, J., & Sibley, C. G. (2007). Right wing authoritarianism, social dominance orientation and the dimensions of generalized prejudice. *European Journal of Personality*, *21*(2), 113–130. https://doi.org/10.1002/per.614

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). 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

Good, J. J., Moss-Racusin, C. A., & Sanchez, D. T. (2012). When do we confront? Perceptions of costs and benefits predict confronting discrimination on behalf of the self and others. *Psychology of Women Quarterly*, *36*(2), 210–226. https://doi.org/10.1177/0361684312440958

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

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*(6), 1318–1333. https://doi.org/10.1002/ejsp.2692

Hurd, N. M., Trawalter, S., Jakubow, A., Johnson, H. E., & Billingsley, J. T. (2022). Online racial discrimination and the role of White bystanders. *American Psychologist*, *77*(1), 39–55. https://doi.org/10.1037/amp0000603

Hyers, L. L. (2007). Resisting prejudice every day: Exploring women's assertive responses to anti-Black racism, anti-Semitism, heterosexism, and sexism. *Sex Roles*, *56*, 1–12. https://doi.org/10.1007/s11199-006-9142-8

Kaiser, C. R., & Miller, C. T. (2001). Stop complaining! The social costs of making attributions to discrimination. *Personality and Social Psychology Bulletin*, *27*(2), 254–263. https://doi.org/10.1177/0146167201272010

Kaiser, C. R., & Miller, C. T. (2003). Derogating the victim: The interpersonal consequences of blaming events on discrimination. *Group Processes & Intergroup Relations*, *6*(3), 227–237. https://doi.org/10.1177/13684302030063001

Kaiser, C. R., & Miller, C. T. (2004). A stress and coping perspective on confronting sexism. *Psychology of Women Quarterly*, *28*(2), 168–178. https://doi.org/10.1111/j.1471-6402.2004.00133.x

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting affective and behavioral responses to racism. *Science*, *323*(5911), 276–278. https://doi.org/10.1126/science.1164951

Lee, M. H. J., Montgomery, J. M., & Lai, C. K. (2024). America's racial framework of superiority and Americanness embedded in natural language. *PNAS Nexus*, *3*(1), Article pgad485. https://doi.org/10.1093/pnasnexus/pgad485

Lee, R. T., Perez, A. D., Boykin, C. M., & Mendoza-Denton, R. (2019). On the prevalence of racial discrimination in the United States. *PLoS One*, *14*(1), Article e0210698. https://doi.org/10.1371/journal.pone.0210698

Li, A. H., Noland, E. S., & Monteith, M. J. (2024). Following prejudiced behavior, confrontation restores local anti-bias social norms. *Personality and Social Psychology Bulletin*. Advance online publication. https://doi.org/10.1177/01461672241229006

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago Face Database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences of the USA*, *116*(20), 9785–9789. https://doi.org/10.1073/pnas.1813486116

Mauduy, M., Priolo, D., Margas, N., & Sénémeaud, C. (2022). When combining injunctive and descriptive norms strengthens the hypocrisy effect: A test in the field of discrimination. *Frontiers in Psychology*, *13*, Article 989599. https://doi.org/10.3389/fpsyg.2022.989599

Meyers, C., Leon, A., & Williams, A. (2020). Aggressive confrontation shapes perceptions and attitudes toward racist content online. *Group Processes & Intergroup Relations*, *23*(6), 845–862. https://doi.org/10.1177/1368430220935974

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Sage Publications.

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

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*, 629–649. https://doi.org/10.1007/s11109-016-9373-5

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

Nelson, J. K., Dunn, K. M., & Paradies, Y. (2011). Bystander anti-racism: A review of the literature. *Analyses of Social Issues and Public Policy*, *11*(1), 263–284. https://doi.org/10.1111/j.1530-2415.2011.01274.x

Ofosu, E. K., Chambers, M. K., Chen, J. M., & Hehman, E. (2019). Same-sex marriage legalization associated with reduced implicit and explicit

antigay bias. *Proceedings of the National Academy of Sciences of the USA*, *116*(18), 8846–8851. https://doi.org/10.1073/pnas.1806000116

Oswald, F., Stevens, S. M., Kruk, M., Murphy, C. I., & Matsick, J. L. (2022). Signaling sizeism: An assessment of body size-based threat and safety cues. *Analyses of Social Issues and Public Policy*, *22*(1), 378–407. https://doi.org/10.1111/asap.12301

Paluck, E. L., & Shepherd, H. (2012). The salience of social referents: A field experiment on collective norms and harassment behavior in a school social network. *Journal of Personality and Social Psychology*, *103*(6), 899–915. https://doi.org/10.1037/a0030015

Park, H. S., & Smith, S. W. (2007). Distinctiveness and influence of subjective norms, personal descriptive and injunctive norms, and societal descriptive and injunctive norms on behavioral intent: A case of two behaviors critical to organ donation. *Human Communication Research*, *33*(2), 194–218. https://doi.org/10.1111/j.1468-2958.2007.00296.x

Perkins, H., & Craig, D. W. (2002). *A multifaceted social norms approach to reduce high-risk drinking: Lessons from Hobart and Williams Smith colleges*. U.S. Department of Education. https://eric.ed.gov/?id=ED470573

Puhl, R. M., & Heuer, C. A. (2009). The stigma of obesity: A review and update. *Obesity*, *17*(5), 941–964. https://doi.org/10.1038/oby.2008.636

Rasinski, H. M., & Czopp, A. M. (2010). The effect of target status on witnesses' reactions to confrontations of bias. *Basic and Applied Social Psychology*, *32*(1), 8–16. https://doi.org/10.1080/01973530903539754

Ratcliff, J. J., Andrus, T., Miller, A. K., Olowu, F., & Capellupo, J. (2023). When potential allies and targets do (and do not) confront anti-Asian prejudice: Reactions to blatant and subtle prejudice during the COVID-19 pandemic. *Journal of Interpersonal Violence*, *38*(23–24), 11890–11913. https://doi.org/10.1177/08862605231188057

Rattan, A., & Dweck, C. S. (2010). Who confronts prejudice? The role of implicit theories in the motivation to confront prejudice. *Psychological Science*, *21*(7), 952–959. https://doi.org/10.1177/0956797610374740

Rimal, R. N., Lapinski, M. K., Cook, R. J., & Real, K. (2005). Moving toward a theory of normative influences: How perceived benefits and similarity moderate the impact of descriptive norms on behaviors. *Journal of Health Communication*, *10*(5), 433–450. https://doi.org/10.1080/10810730591009880

Sanchez, D. T., Chaney, K. E., Manuel, S. K., Wilton, L. S., & Remedios, J. D. (2017). Stigma by prejudice transfer: Racism threatens White women and sexism threatens men of color. *Psychological Science*, *28*(4), 445–461. https://doi.org/10.1177/0956797616686218

Sanchez, D. T., Himmelstein, M. S., Young, D. M., Albuja, A. F., & Garcia, J. A. (2016). Confronting as autonomy promotion: Speaking up against discrimination and psychological well-being in racial minorities. *Journal of Health Psychology*, *21*(9), 1999–2007. https://doi.org/10.1177/1359105315569619

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

Sechrist, G. B. (2010). Making attributions to and plans to confront gender discrimination: The role of optimism. *Journal of Applied Social Psychology*, *40*(7), 1678–1707. https://doi.org/10.1111/j.1559-1816.2010.00635.x

Shelton, J. N., Richeson, J. A., Salvatore, J., & Hill, D. M. (2006). Silence is not golden: The intrapersonal consequences of not confronting prejudice. In S. Levin & C. van Laar (Eds.), *Stigma and group inequality* (pp. 79–96). Psychology Press. https://doi.org/10.4324/9781410617057

Sidanius, J., & Pratto, F. (2001). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.

Siegel, A. A., & Badaan, V. (2020). # No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, *114*(3), 837–855. https://doi.org/10.1017/S0003055420000283

Sommers, S. R., & Norton, M. I. (2006). Lay theories about White racists: What constitutes racism (and what doesn't). *Group Processes & Intergroup Relations*, *9*(1), 117–138. https://doi.org/10.1177/1368430206059881

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

Tankard, M. E., & Paluck, E. L. (2017). The effect of a Supreme Court decision regarding gay marriage on social norms and personal attitudes. *Psychological Science*, *28*(9), 1334–1344. https://doi.org/10.1177/0956797617709594

Thai, M., & Nylund, J. L. (2024). What are they in it for? Marginalised group members' perceptions of allies differ depending on the costs and rewards associated with their allyship. *British Journal of Social Psychology*, *63*(1), 131–152. https://doi.org/10.1111/bjso.12670

Viscusi, W. K., Huber, J., & Bell, J. (2011). Promoting recycling: Private values, social norms, and economic incentives. *American Economic Review: Papers and Proceedings*, *101*(3), 65–70. https://doi.org/10.1257/aer.101.3.65

Wessel, J. L., Lemay, E. P., Jr., & Barth, S. E. (2023). You(r behaviors) are racist: Responses to prejudice confrontations depend on confrontation focus. *Journal of Business and Psychology*, *38*(1), 109–134. https://doi.org/10.1007/s10869-022-09811-5

Williams, M. J., & Eberhardt, J. L. (2008). Biological conceptions of race and the motivation to cross racial boundaries. *Journal of Personality and Social Psychology*, *94*(6), 1033. https://doi.org/10.1037/0022-3514.94.6.1033

Woodzicka, J. A., & LaFrance, M. (2001). Real versus imagined gender harassment. *Journal of Social Issues*, *57*(1), 15–30. https://doi.org/10.1111/0022-4537.00199

Zapata, J., Sulik, J., von Wulffen, C., & Deroy, O. (2024). Bystanders' collective responses set the norm against hate speech. *Humanities and Social Sciences Communications*, *11*(1), 1–13. https://doi.org/10.1057/s41599-024-02761-8

Zitek, E. M., & Hebl, M. R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology*, *43*(6), 867–876. https://doi.org/10.1016/j.jesp.2006.10.010