THE ROLE OF TRUST IN REDUCING CONFRONTATION-RELATED SOCIAL COSTS

by

Laura Hildebrand

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Psychological Sciences

West Lafayette, Indiana

August 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Margo J. Monteith, Chair

Department of Psychological Sciences

Dr. Kipling Williams

Department of Psychological Sciences

Dr. Ximena B. Arriaga

Department of Psychological Sciences

Dr. Yk Hei F. Kung

Department of Psychological Sciences

Approved by:

Dr. Kimberly P. Kinzig

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Interpersonal confrontation of prejudiced remarks or behavior consistently reduces people's subsequent expressions of bias but comes with social costs (e.g., dislike, avoidance) for the confronter. However, research has yet to address how interpersonal dynamics influence reactions to confrontations. The present research (N = 1,019) integrated the close relationships and prejudice reduction literatures to investigate whether interpersonal trust reduces confrontation's social costs. Study 1 provided correlational evidence that trust mitigated the effect of sexism confrontation on negative other-directed affect (e.g., irritation), and in turn, social costs. Manipulation of confrontation and trust in Study 2 provided causal evidence that trust buffered against social costs: Participants confronted about racism directed more negative other-directed affect and social costs at their study partner than not-confronted participants; however, these effects were mitigated among participants who underwent a trust-building exercise with their confronter. Study 3 showed that the effect of trust on social costs extended to an ecologically-valid context. Participants were confronted about racism by either a friend or stranger. Social costs were buffered for people confronted by friends versus strangers, and this effect was serially mediated by pre-existing trust and negative other-directed affect. Importantly, confrontation reduced subsequent expression of bias in all studies. Practically, findings provide a reassuring message to people who do not confront due to fear of social costs: To the extent that interpersonal trust can be created or is pre-existing, social costs should be mitigated. Theoretically, the present research highlights how insights from close relationships research can advance our understanding of prejudice reduction.

INTRODUCTION

Interpersonal confrontations are a powerful prejudice reduction strategy. Imagine, for instance, someone who confronts a coworker for a prejudiced statement. That confrontation will reduce the coworker's future expressions of prejudice (Chaney & Sanchez, 2018; Czopp et al., 2006; Parker et al., 2018), convey that prejudice does not belong in that environment (Koudenburg et al., 2020; Moser & Branscombe, 2021; Monteith et al., in press), and promote feelings of belonging and safety among marginalized individuals (Hildebrand et al., 2020). Despite this myriad of benefits, confrontations also come with a dark side: The person who confronted (i.e., the confronter) will experience backlash. The coworker, as well as others who witness the confrontation, will derogate, dislike, and even distance themselves from the confronter (Czopp & Monteith, 2003; Czopp et al., 2006; Monteith et al., in prep; Parker et al., 2018).

But what if the confrontee (i.e., the person who was confronted) was more than a coworker? What if the confrontee was a friend, or some kind of close other? How might relationship dynamics influence the interpersonal backlash that confronters typically experience following a confrontation? To date, existing research has yet to consider the interpersonal dynamics of bias confrontations, even though such dynamics may influence confrontation outcomes. The present research begins to fill this gap by examining how trust influences interpersonal outcomes following a confrontation. Does trust reduce negative affect directed at the confronter and, in turn, interpersonal costs? The present research uniquely draws upon both confrontation and close relationships research to answer this question.

The Social Costs of Confronting

A substantial body of research shows that bias confrontations (i.e., calling someone out for biased statements or behavior) reduce subsequent expressions of bias (Chaney & Sanchez, 2018; Czopp et al., 2006; Gulker et al., 2013; Mallet & Wagner, 2011; Parker, et al., 2018; Rasinski & Czopp, 2010; for a review, see Monteith et al., in press). Certain factors make a bias confrontation more or less effective (e.g., whether the confrontation is presented with evidence, whether the confrontation is delivered by a target-group or dominant-group member; Czopp & Monteith, 2003; Gardner & Ryan, 2020; Gulker et al., 2013; Parker et al., 2018; Rasinski & Czopp, 2010; Thai et

al., 2021). Nevertheless, the bias-reducing effect of confrontation is remarkably consistent: Compared to no confrontation, confrontation reduces bias, regardless of confrontation style (Becker & Barreto, 2014; Burns & Monteith, 2019; Czopp et al., 2006; Parker et al., 2018), the type of "ism" targeted by the confrontation (e.g., racism vs. sexism; Burns & Monteith, 2019; Parker et al., 2018), or whether a target- or dominant-group member confronts (Czopp et al., 2006). Furthermore, the bias-reducing effect of confrontations persists up to a week later (Chaney & Sanchez, 2018; Monteith, Hildebrand, & Mallett, in prep). Thus, bias confrontations are an excellent strategy for those wishing to curb bias.

Yet, bias confrontations also come with social costs for the confronter. Both bystanders and the confrontee (i.e., the target of the confrontation) evaluate the confronter more negatively (Czopp et al., 2006; Monteith et al., in prep; Parker et al., 2018). For instance, across two studies conducted by Czopp and colleagues (2006), participants believed that they were interacting with another participant over the computer; in reality, however, they were interacting with a confederate. After completing a task designed to elicit a stereotypic response, the confederate either confronted or did not confront the participant about using racial stereotypes. Afterwards, confronted participants disliked the confederate more and evaluated the confederate more negatively than not-confronted participants. Across two studies, Monteith and colleagues (in prep) demonstrated negative other-directed affect, such as irritation and annoyance with the confronter, mediate this effect. Specifically, participant felt more annoyed and irritated by the confrontation, which in turn was associated with greater social costs. These social costs are particularly strong for target-group confronters (Becker & Barreto, 2014; Czopp & Monteith, 2003; Czopp et al., 2006; Drury & Kaiser, 2014; Elizier & Major, 2012; Gervais & Hillard, 2014; Gulker et al., 2013; Rasinki & Czopp, 2010; Schultz & Maddox, 2013; though see Mallett & Wagner, 2011) and for those confronting sexism versus racism (Czopp & Monteith, 2003; Woodzicka et al., 2015).

Confrontation-related social costs even persist up to a week after the confrontation (Monteith et al., in prep). In one study, participants were either confronted or not confronted by a confederate for racial bias. Participants then reported negative other-directed affect and social costs towards the confronter. Five-to-seven days later, participants were contacted for an online follow-up, in which they again rated the confederate. In line with other research, confronted participants reported more negative other-directed affect and social costs than not-confronted participants. At

the five-to-seven-day follow-up, these social costs had not dissipated and in fact had increased compared to the initial assessment. Thus social costs are lasting problem for confronters.

People have an innate need to be accepted by others (Baumeister & Leary, 1995), and social costs hurt the confronter by threatening that need. Furthermore, social costs make people hesitant to confront instances of bias (Kawakami et al., 2019; Shelton & Stewart, 2004; Woodzicka & LaFrance, 2001). For instance, in a study by Shelton and Stewart (2004), female participants were interviewed for a mock job position by a man who asked sexist questions (e.g., "Do you think it is important for women to wear bras to work?"). Before the interview participants were informed that the job was either a prestigious, high-salary, and highly sought-after position or that the job was a low-salaried position at a charity, with little competition. Thus, the potential costs of confronting (i.e., being disliked and losing out on the job) were higher for women who interviewed for the more versus less prestigious job. Results revealed that women who interviewed for the more prestigious job were less likely to confront than women who interviewed for the less prestigious job. In other words, when the negative consequences of confronting were higher versus lower, participants were less likely to confront.

People who fail to confront often experience guilt, regret, and rumination (Shelton et al., 2006). Furthermore, people who are expected to confront (e.g., someone who has expressed a commitment to egalitarianism) but do not are, ironically, negatively evaluated by others (Becker & Barreto, 2014; Czopp, 2013). Most importantly, failures to confront allow the biased behavior to continue and perpetuate social norms allowing bias (Czopp, 2013; Mallett et al., 2019; see also Blanchard et al., 1994).

Overall, existing research indicates that social costs hurt the confronter and reduce the likelihood of confrontation. How then can one alleviate confrontation-related social costs?

Does the Confrontee Trust the Confronter?

Trust may be one answer. Specifically, the extent to which the confrontee directs social costs towards the confronter may depend upon the extent to which the confrontee trusts the confronter. Trust has been conceptualized in a variety of ways. For instance, in 1967, Rotter defined trust as the trait-level belief that other people can be relied on. In the present research, we focus on trust as conceptualized in the close relationships literature. Early research in this domain defined trust as the "confidence that one will find what is desired from another, rather than what

is feared" (Deutsch, 1973). Although there is some variation in contemporary definitions of trust (Righetti & Finkenauer, 2011; Rousseau et al., 1998; Simpson, 2007), it is usually defined in terms of one's willingness to be vulnerable with another person based on positive expectations that the trusted person cares about and will act in ways that benefit oneself (Murray et al., 2006; Murray et al., 2011; Rempel et al., 1985; Rousseau et al., 1998).

Importantly, the present research focuses on trust, not as a stable individual difference, but rather, as a psychological state that depends on the specific relationship between oneself and the trusted (or not-so-trusted) person (Rusbult & Van Lang, 2008; Simpson, 2007). In the context of confrontations, the relevant relationship is between the confronter and the confrontee, and we are concerned with how much the confrontee trusts the confronter (rather than the other way around). Does the confrontee trust that the confronter "has their back" and their bests interests at heart?

According to the risk regulation model, trust buffers against rejection (e.g., criticism, conflict) from close others (Campbell et al., 2010; Murray et al., 2003; Murray et al., 2012; Rempel et al., 1985; Shallcross & Simpson, 2012). For instance, in one study, participants were led to believe that their romantic partner criticized them (i.e., generated a long list of faults about them; Murray et al., 2012). Participants showed more resilience to that criticism (as measured by physiological challenge versus threat responses) the more they trusted their partner (Murray et al., 2012). In related research, trust in their romantic partner reduced the extent to which participants distanced themselves following criticism from the romantic partner (Murray et al., 2011). People who explicitly trusted their partner were even more supportive and accommodating with their partner in a relationship conflict (Shallcross & Simpson, 2012).

Trust even influences the way individuals construe psychologically threatening situations. Compared to people with less trust towards their romantic partner, people with more trust towards their romantic partners perceived their partners as more accommodating during a difficult discussion (Shallcross & Simpson, 2012), associated their partner with positive traits more quickly and negative traits more slowly (Murray et al., 2011), perceived their partner in a more positive light (Murray et al., 2000; Rempel et al., 2001), and even remembered their partner's past transgressions as less severe, less frequent, and less consequential (Luchies et al., 2013). These effects of trust extend beyond romantic partners to ingroup members (Cruwys et al., 2021) and other close relationships, such as friends (Monsor, 1992; Yoo et al., 2011). Taken together, this

research indicates that trust reduces the "sting" of rejection and leads to constructive versus destructive behaviors.

Although existing research has yet to examine trust in the context of bias confrontations, trust should play the same ameliorating role. In the same way that trust buffers against the effect of criticism and rejection (e.g., Murray et al., 2006; Murray et al., 2012), trust may buffer against the perceived criticism, and subsequent social costs, associated with confrontation. Of course, people do not like to receive criticism, so even confrontees who highly trust the confronter may still feel some hurt. However, compared to confrontees with less trust towards the confronter, confrontees with more trust may perceive the confrontation more positively.

Why May Trust Reduce Social Costs?

Why may trust buffer against confrontation-related social costs? When the confrontee trusts that the confronter has their best interests at heart, they may be less likely to interpret the confrontation as a threat or rejection. In other words, as trust increases, people may interpret the confrontation more benevolently. This benevolent interpretation may reduce confrontation-related social costs in two possible ways. First, trust may reduce the extent to which people interpret the confrontation as threatening their nonprejudiced image. Existing research indicates that people desire to maintain a nonprejudiced image in the eyes of others. For instance, White people's overwhelming concern in race-related situations (e.g., interracial interactions; taking a race-related IAT) is that they will be viewed as prejudiced (Bergsieker, Shelton, & Richeson, 2010; Frantz et al., 2004; Vorauer, Main, & O'Connell, 1998; Vonach, Reynolds, Winegard, & Baumeister, 2018). Findings that people change their behavior to align with nonbiased social norms also suggests that people are concerned about being seen as prejudiced (Crandall et al., 2002; Monteith et al., 1996; Murrar et al., 2020).

People may react negatively to confrontations because they perceive the confrontation as threatening meta-perceptions of that image. In support of this possibility, research indicates that people often respond to confrontations by trying to bolster their nonprejudiced image in the eyes of the confronter. In three studies by Czopp and colleagues (2006), participants had the opportunity to respond to the confronter after being confronted about stereotypic responses. Coders later categorized these responses according to whether participants denied being influenced by prejudice. Approximately half of the responses involved denial: Participants denied that race

played a role in their responses (e.g., "I wasn't looking at [race in] the pictures") or that they were a prejudiced person (e.g., "I'm not racist"). These denials suggest that the participant was concerned with the confronter viewing them as a prejudiced person.

Trust may reduce confrontation-related social costs by alleviating this threat. Specifically, as trust increases, a confrontee may be less likely to interpret the confrontation as a threat to their nonprejudiced image. This possibility is in line with the idea that trusted individuals will see the best in oneself, even in the face of flaws and other shortcomings (Murray et al., 2006; Murray et al., 2012; Simpson, 2007; Shallcross & Simpson, 2012). In short, with a trusted confronter, the confrontee may be less likely to perceive the confrontation as a threat or rejection; the confrontee may believe that the confronter sees the best in them and is not viewing them as inherently bigoted.

A second possible way in which trust may alleviate social costs is by reducing the negative affect the confrontee feels toward the confronter (i.e., negative other-directed affect). Existing research demonstrates that negative other-directed affect mediates the relation between confrontation and social costs (Monteith et al., in prep). Trust may buffer against confrontation-related social costs by reducing such annoyance and irritation. This possibility is consistent with the meaning-making literature, which shows that people who cognitively reappraise a stressful event experience less emotional distress than people who do not reappraise the event (Park et al., 2010). Similarly, people who trust the confronter should reappraise the confrontation more positively and subsequently feel less negative other-directed affect. This possibility is also consistent with research indicating that trust leads people to engage in more constructive versus destructive relationship behaviors (e.g., Luchies et al., 2013; Murray et al., 2011; Murray et al., 2000; Rempel et al., 2001; Shallcross & Simpson, 2012). Annoyance and irritation, as two destructive relationship behaviors, may be similarly inhibited by trust.

Overview of Proposed Research

Does trust reduce the social costs (e.g., negative impressions and evaluations of the confronter) elicited by confrontation? What mechanism explains the trust-induced reduction in social costs? The proposed research aims to answer these questions by examining the effect of trust, negative other-directed affect, and nonprejudiced image threat on people's interpersonal reactions to confrontations.

Our first study aimed to replicate confrontation research and, more importantly, examine the relations among trust, social costs, negative other-directed affect, and image threat. Study 2 then (a) tested the causal effect of trust on social costs and (b) further explored the mediating role of negative other-directed affect and nonprejudiced image threat. Finally, Study 3 examined these research questions within an ecologically-valid context. Specifically, we examined the difference in confrontation-related social costs among dyads who naturally differ in trust (i.e., people who believed they were confronted by friends versus strangers).

Overall, these three studies are the first to our knowledge to examine how trust influences interpersonal outcomes following a confrontation. Although confrontation often occurs in a dyad, little existing research has examined confrontation from a close relationship perspective. The novel integration of these two research areas provides theoretical and empirical insight to answer the practically-important question of how to reduce the social costs associated with confrontation.

STUDY 1[1]

Is trust associated with social costs? What explains the potential relation between trust and social costs? Study 1 aimed to answer these research questions by examining (a) whether trust was negatively related to social costs, negative other-directed affect, and nonprejudiced image threat within the confrontation context and (b) whether negative other-directed affect and image threat mediate the relation between trust and social costs. As outlined above, trust may reduce social costs by reducing the nonprejudiced image threat typically associated with confrontation; alternatively, trust may reduce social costs by reducing the annoyance and irritation also associated with confrontation. By conducting exploratory mediation models, we aimed to gain more understanding about the mechanisms behind the relation between trust and social costs.

Finally, we also tested whether, replicating past research, the confrontation increased negative self- and other-directed affect, increased social costs, and decreased bias. Note that Study 1 originally included two variations of a confrontation condition: Participants received either a confrontation only, or a confrontation that included a statement designed to alleviate image threat (i.e., "I'm sure you're NOT sexist or anything like that and that it was just a mistake"). Analyses of the image threat manipulation check indicated that participants in the image-protection and confrontation-only conditions reported comparable levels of image threat, $t(319) = .77$, $SE = .12$, $p = .43$, 95%CI[-.33, .14], $d = .11$ (see Appendix A for details). Given the manipulation was ineffective, we collapsed across the two confrontation conditions in analyses reported in the main text of this manuscript. Our preregistration of Study 1, https://osf.io/y4wz2/?view_only=959eadeaeace49f484e07dd02f649ba5, references the image-threat condition. Given the failed manipulation check, the portions of the pre-registration that focus on this manipulation are no longer relevant. In addition, our analyses occasionally depart from the pre-registration, which is noted where relevant.

For this and all studies, materials, data, syntax, and output can be accessed at https://osf.io/vkrp3/?view_only=3b468c8bbe1b4366bc181e1378f12021.

[1] This study was originally proposed as the third dissertation study. However, after considering the results and the overall research questions, we decided to present it first in this program of research.

Design and Participants

Participants were randomly assigned to either a no-confrontation or confrontation condition (which collapsed across confrontation-only and image-protection confrontation conditions, as explained above).

We collected data from 354 undergraduate students for a study on "behavior in workplace settings." Because the present research examines stereotypes that may be specific to US culture, in this and all studies, only people born in the United States were allowed to participate. Data were excluded from four participants (all in the confrontation condition) who did not give post-session consent to analyze their data. In addition, two researchers independently read participant responses to funnel debriefing questions and identified responses in which participants suspected they were not interacting with a real person. The researchers discussed these cases and jointly made final decisions about suspicion. Based on an *a priori*, pre-registered decision, 28 suspicious participants (three who were not confronted, 25 who were confronted) were excluded from analyses, leaving a final sample of 322 participants (78% White, 10.6% Asian, 4.0% Hispanic or Latinx, 7.4% other; 176 women, 140 men, 4 non-binary, 1 transwoman, 1 transman). A sensitivity analysis specifying a two-tailed independent samples t-test, with an alpha of .05, indicated that the present research had 80% power to detect small-to-medium effect sizes (d = .33).

Procedure

Sessions included between two and ten participants for this in-person study. Upon arrival, participants were led to small, individual computer rooms, where they learned that they would be interacting with another participant over the computer. However, all partner responses were pre-programmed and delivered by the computer; the participant did not actually interact with anyone over the computer. In sessions with an odd number of participants, a research assistant posed as a participant.

Participants first completed a short profile about themselves (e.g., age, favorite TV show), which was ostensibly sent to their study partner. In return, they received their study partner's supposed responses to the same profile. This served as a filler task to boost the cover story and reduce suspicion. To avoid potential assumptions about confronter gender, the profile purposefully did not include a name or gender information.

Then participants completed a moral decision-making task, adapted from Mallett and Wagner (2011), with their study partner. Specifically, participants described how they would respond to three moral dilemmas. Those responses were sent to their study partner to read. Their study partner supposedly described their response to three different moral dilemmas, which were sent to the participant to read.

Critically, all three of the participant's moral dilemmas featured stereotypically masculine jobs (i.e., CEO, computer programmer, and surgeon; U.S. Bureau of Labor Statistics, 2021), which should activate gender role schemas and create a situation where participants respond with gender-exclusive (e.g., "he") versus gender-neutral (e.g., "they," "he or she") language (See Appendix B for the full dilemmas). Approximately half of participants (49.7%) used at least one gender-exclusive pronoun: 12.7% used one gender-exclusive pronoun, 11.8% used two; and 25.5% used three or more. This is contrary to past research, in which 80% of participants used a gender-exclusive pronoun (Mallett & Wagner, 2011), However, past research indicates that all (Mallett & Wagner, 2011) or almost all (98.8%; Monteith, Hildebrand, & Mallett, in prep) of participants remember using gender-exclusive pronouns, even when they did not actually use such pronouns. Excluding participants who did not use a gender-exclusive pronouns did not change the results, so we included them to maximize power.

After the moral decision-making task, participants messaged their study partner with any reactions they had during the task. The confrontation was embedded within the study partner's message to the participant. In the *confrontation* condition, participants read:

> I thought the task went okay, some of those dilemmas were tricky to respond to. But I noticed that for certain dilemmas you used "he" to refer to the person. Are you assuming the computer scientist, the CEO, the surgeon is a man? Women can have jobs like that too. I just think we just need to be careful not to make assumptions about gender.

In the *no-confrontation condition*, participants simply read the first sentence (i.e., "I thought the task went okay…"). Because the interaction was pre-programmed, whether or not the participant received the confrontation did not depend on their pronoun use.

After the confrontation, participants completed measures and demographic questions. Finally, participants were probed for suspicion and debriefed.

Measures

All measures other than the Sexist Language Detection Task were completed on a 1 (*strongly disagree)* to 7 (*strongly agree*) scale. See Appendix C for Study 1 measures.

Nonprejudiced Image Threat

Participants responded to 18 trait items assessing their meta-perceptions of "how the study partner thinks about" them. Embedded within 14 filler items were four critical items developed for this study that assess the extent to which confrontation threatens the participant's nonprejudiced image (e.g., "My study partner thinks of me as fair-minded" [reverse-scored], "My study partner thinks of me as unbiased" [reverse-scored]). A principal components analysis using varimax rotation indicated that all items loaded onto 1 factor that explained 73.15% of variance (loadings > .77).

Affect

Participants reported how they felt in the current moment using 17 affect items. In line with past research (e.g., Monteith, 1993), we formed two primary affect indices by averaging the relevant items: Negative self-directed affect (*negself*; six items; e.g., "guilty," "self-critical") and negative other-directed affect (*negother*; three items; e.g., "irritated," "threatened"). Of lesser importance, positive affect (*positive;* α = .83; four items; e.g., "happy," "optimistic") and discomfort (*discomfort*; α = .77; three items; e.g., "uncomfortable," "tense") indices were also formed (see Appendix D for analyses involving positive affect and discomfort). The item "proud" did not reliably load onto any indices and was not examined further.

Social Costs

Social costs were assessed with partner impressions (Czopp et al., 2006; α = .91; nine items; e.g., "I probably wouldn't be friends with someone like the other participant"), evaluations of the interaction (Mallett & Wagner, 2011; α = .91; seven items; e.g., "I enjoyed working on the task with the other participant"), and desire to avoid future contact (α = .79; three items; e.g., "I don't

want to interact with my study partner again").[2] Results were redundant across measures. So, based on an *a priori*, pre-registered decision, each measure was standardized and scores were averaged to create a single *social costs* index. Higher scores indicate greater social costs.

Trust

Trust was assessed with three items from the Faith dimension of the Trust scale (Rempel et al., 1985; e.g., "Even if I do or say something flawed, I feel like I can rely on my study partner to see me in a positive way"), and three self-created items (e.g., "My study partner seems to have doubts about whether or not we are compatible"). A principal components analyses with varimax rotation suggested a two-factor solution that explained 72.89% of the variance (loadings > .62). Nevertheless, in line with the pre-registered measurement plan, we chose to average all items to form a one-factor *trust* solution that explained 54.45% of the variance.

Sexist Language Detection

Participants' biased responding was assessed through the Sexist Language Detection Task (Mallett & Wagner, 2011; McMinn, Williams, & McMinn, 1994; Swim, Mallett, & Stangor, 2004). Specifically, participants worked as quickly as possible to identify "every writing problem you find, including problems with grammar, spelling, punctuation, and sexist and otherwise discriminatory language" in 30 sentences that were supposedly being pilot tested for a separate study (see Appendix C for the full task). For instance, the sentence, "The most recently hired secretary was asked to check her boses mail twice a day" contained one spelling error ("boses") and one instance of sexist language (the assumption that the secretary is a woman). Four experimenters coded for detection of spelling/grammar errors (ICC = .99) and sexist language (ICC = .99). For this and all coding across studies, the researchers did not know participants' experimental condition while coding. Ratings were averaged across coders. We then calculated a *sexist language detection score* by adding the number of sexist language errors detected.

[2] In all studies, we pre-registered a feeling thermometer assessing how warm/cold participants felt toward their partner as part of the social costs composite. However, we later decided to exclude it from the social costs measure due to its construct overlap with the affect measure. Results did not change when feeling thermometer was included.

Results

How Does Confrontation Influence Measures?

Each dependent variable was predicted using an independent t-test (no-confrontation vs. confrontation). When Levene's Test for Equality of Variances was violated, we report analyses where equal variances are not assumed. For this reason, degrees of freedom occasionally differ. Confidence intervals are for the mean differences between the no-confrontation and confrontation conditions. See Table 1 for reliability, descriptive statistics, and inter-measure correlations. See Table 2 for descriptive statistics as a function of condition.

Table 1. Descriptive Statistics, Reliability, and Inter-Measure Correlations, Study 1

	α	*M* (*SD*)	1)	2)	3)	4)	5)
1.) NegSelf	0.91	2.42 (1.34)	—	—	—	—	—
2.) NegOther	0.82	2.18 (1.34)	.35***	—	—	—	—
3.) Social Costs	0.94	-.02 (0.88)	.15**	.68***	—	—	—
4.) Trust	0.83	4.02 (0.96)	-.24***	-.56***	-.76***	—	—
5.) Nonprejudiced Image Threat	0.80	3.87 (1.13)	.35***	.48***	.61***	-.64***	—
6.) Sexist Language Detection		7.76 (5.13)	.21***	.11	.13*	-.15**	-.22***

Note. For social costs, the reliability is for the linear composite (Nunnally, 1978).

*$p < .05$. **$p < .01$. **$p < .001$.

Table 2. Measures as a Function of Confrontation Condition, Study 1

	No Confrontation (n = 116)	Confrontation (n = 206)
Negself	1.69 (0.79)	2.84 (1.40)
Negother	1.32 (0.75)	2.66 (1.36)
Social Costs	-.67 (0.56)	0.35 (0.81)
Trust	4.60 (0.73)	3.67 (0.90)
Nonprejudiced Image Threat	2.90 (0.80)	4.42 (0.90)
Sexist Language Detection	5.64 (4.66)	8.95 (5.01)

Note. Numbers in parentheses are cell standard deviations. All cells significantly differed at $p < .001$.

We first tested whether the confrontation paradigm used in this research had the intended outcomes. Specifically, did the confrontation increase negative self- and other-directed affect, increase social costs, and decrease bias? As expected, compared to participants who were not confronted, confronted participants reported more negative self-directed affect, $t(320) = 9.48$, $p < .001$, 95%CI [-1.43, -.88], $d = 1.19$, more negative other-directed affect, $t(320) = 11.45$, $p < .001$, 95%CI[-1.58, -.1.11], $d = 1.14$, and greater social costs, $t(307) = 13.16$, $p < .001$, 95%CI[-1.17, -.86], $d = 1.38$. Confronted participants detected more sexist language than participants in the no-confrontation condition, $t(320) = 5.85$, $p < .001$, 95%CI[-4.43, -2.20], $d = .68$. Thus, the present findings replicated past research (e.g., Chaney & Sanchez, 2018; Czopp et al., 2006; Gulker et al., 2013; Mallett & Wagner, 2011; Parker et al., 2018) for all measures. However, contrary to past research, negself did not mediate the relation between confrontation condition and sexist language detection, $B = .45$, $SE = .28$, 95%CI[-.09, 1.01].

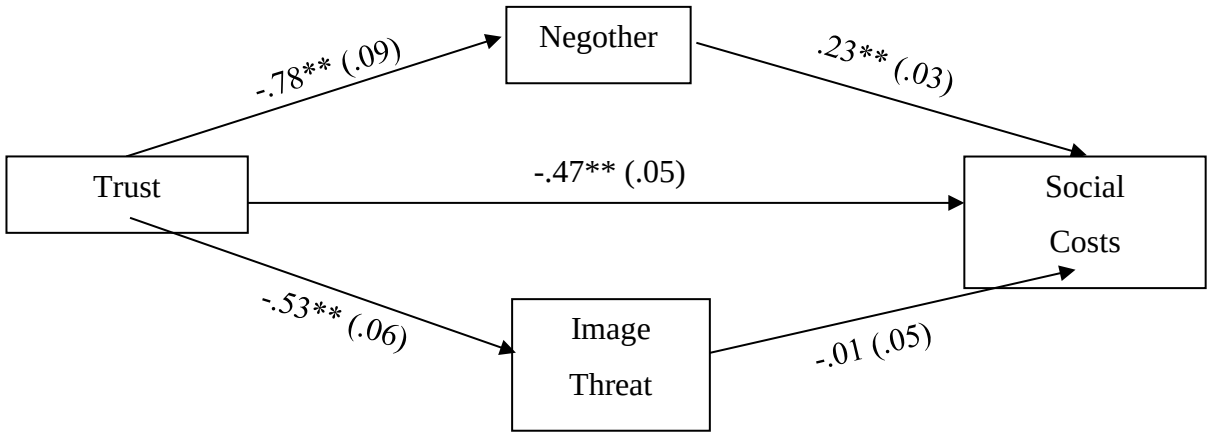
New to the present research, we examined whether the confrontation increased image threat and decreased trust. As expected, confronted participants reported more image threat, $t(320) = 15.22$, $p < .001$, 95%CI[-1.72, -1.33], $d = 1.78$, and less trust, $t(281) = 10.77$, $p < .001$, 95%CI[.81, 1.17], $d = 1.18$, than participants in the no-confrontation condition. This finding is in line with the idea that confrontation is a form of criticism that influences interpersonal dynamics between the confronter and the confrontee.

How Does Trust Influence Interpersonal Dynamics Within the Confrontation Context?

We then examined (a) the relations between trust and the interpersonal measures (i.e., social costs, negative other-directed affect, and image threat) and (b) the structure of those relations. Given our interest in trust within the confrontation context, only confronted participants (n = 206) were included in the following analyses. Note that these analyses were not preregistered.

Trust was strongly correlated with social costs, r = -.72, p < .001, negative other-directed affect, r = -.51, p < .001, and image threat, r = -.53, p < .001. Thus, as trust in the confronter increased, participants reported less dislike of the confronter, less annoyance and irritation, and less threat to their nonprejudiced image. These results provide initial correlational evidence regarding the potential role of trust in reducing the negative interpersonal consequences associated with confrontation.

Why is trust negatively related to social costs? We considered it possible trust may mitigate social costs by reducing the image threat caused by confrontation; alternatively, trust may mitigate social costs by reducing the negative other-directed affect also caused by confrontation. To test these possibilities, we compared negative other-directed affect and image threat as mediators for the relation between trust and social costs. We conducted a simultaneous mediation analysis using Hayes's (2018) PROCESS (V3; Model 4; 5000 bootstraps). We entered trust as the predictor, social costs as the outcome, and negother and image threat as simultaneous mediators (see Figure 1). Negother was a significant mediator, B = -.18, SE = .03, 95%CI[-.25, -.12], but image threat was not, B = .008, SE = .03, 95%CI[-.04, .05]. Thus, within the confrontation context, trust was associated with less annoyance and irritation following confrontation, which in turn was associated with fewer social costs. In contrast, image threat was not a significant mediator. Overall, these results provide initial evidence that negother, not image threat, may be responsible for the negative association between trust and confrontation-related social costs.



Note. Path values are unstandardized coefficients with standard errors presented in parentheses. ** $p < .001$.

Figure 1. Simultaneous moderated mediation of negother and image threat on the relation among confrontation, trust, and social costs.

Discussion

Study 1 is the first to our knowledge to consider the role of interpersonal factors (i.e., trust) within the confrontation context. In addition to replicating aspects of past confrontation research, Study 1 provides evidence that, within the confrontation context, trust is associated with fewer social costs.

Study 1 also provides the first test of two possible underlying processes in the trust-social costs relation: Negative other-directed affect and image threat. Although trust was also negatively associated with both negative other-directed affect and image threat, only negative other-directed mediated the relation between trust and social costs. Thus, trust may buffer against social costs, not because it reduces meta-perceptions of prejudice, but rather because it reduces feelings of irritation and annoyance with the confronter. In Study 2 we test the replicability of these findings by further examining the potential mediating roles of both negative other-directed affect and image threat.

It is worth noting that, in the present study, trust was only assessed after the confrontation, rather than before. Past research on trust has primarily focused on pre-existing trust (i.e., trust before the rejection or criticism), as the criticism itself may influence trust. Indeed, in the present research, confronted participants reported less trust in their study partner than not-confronted

participants. Thus, in Study 1, trust was not a "pure," unadulterated variable but rather was influenced by the criticism that came before. Despite such variance, the present research still finds support for the relations among trust, negative other-directed affect, and social costs. Nevertheless, in Study 2 we focus on pre-existing trust (i.e., trust that comes before the confrontation).

Overall, Study 1 provides initial evidence that trust may ameliorate the negative interpersonal consequences associated with confrontation. Yet, such findings are correlational, and it is important to determine whether trust is the key psychological variable in the mitigation of social costs. In Study 2, we move beyond this correlational evidence to show the causal effect of trust on confrontation-related social costs.

STUDY 2

In Study 2, participants were randomly assigned to complete either a dilemma-solving task that built trust with their study partner (the *trust-present* condition) or a *trust-neutral* interaction task. We expected that confrontation and trust would interact when predicting interpersonal outcomes. Specifically, we hypothesized that, although confronted participants would evaluate the confronter more negatively than non-confronted participants, these effects would be attenuated among participants in the trust-present versus trust-neutral condition.

Does negative other-directed affect or nonprejudiced image threat mediate the relation between trust and social costs? Study 2 continued to explore the potential mediating roles of negative other-directed affect and nonprejudiced image threat. Specifically, Study 2 tested whether the Study 1 mediation results would replicate when (a) trust was manipulated rather than measured and (b) a different type of bias was examined. As in Study 1, will negative other-directed affect explain the reduction in social costs among confronted participants in the trust-present condition? Or will image threat serve as a mediator instead?

Method

Design and Participants

The study used a 2 (confrontation: bias confrontation vs. no confrontation) x 2 (trust: trust-present vs. trust-neutral) design. When deciding the anticipated effect size, we considered that (a) confrontation research routinely produces medium-sized effects (d = .5; e.g., Czopp et al., 2006) and (b) a pilot study on the effect of the trust manipulation yielded a large simple effect (d = .87). However, the pilot study only included one predictor (trust), while the present study examined the interaction between two predictors. Given these considerations, we conducted an *a priori* power analysis in GPower (Faul et al., 2009), specifying ANOVA (fixed effects, main effects, and interactions), a small-to-medium effect size (d = .32 or f = .16), and an alpha of .05, which indicated that a minimum of 309 participants (or approximately 77 participants per cell) was needed for 80% power. Given the planned moderated mediation analyses and possible participant exclusions (e.g.,

suspicious participants, participants who didn't use stereotypes before the confrontation), we increased our target sample size to 375 participants.

We recruited 365 non-Black domestic undergraduate students for a study on "the factors that influence group/dyad communication." Participants received partial course credit for participating. Based on *a priori*, pre-registered decisions, data were excluded from 26 participants who did not respond stereotypically prior to the confrontation (14 in the confrontation condition; 12 in the no-confrontation condition), four participants who did not give post-session consent to use their data (two in the confrontation condition; two in the no-confrontation condition), three participants who missed two or more attention checks, one participant provided the same response to every item, and one participant who self-reported autism and "no ability to read social cues."

Two researchers identified and discussed responses in which participants expressed suspicion about their study partner. The researchers discussed these cases and jointly made final decisions about suspicion. Based on an *a priori*, pre-registered decision, 18 suspicious participants (17 in the confrontation condition; 1 in the no-confrontation condition) were excluded from analyses. The final sample included 312 participants (80.8% White, 10.6% Asian, 5.1% Hispanic/Latinx, 3.5% other; 175 women, 134 men, 2 non-binary, and 1 pangender).

The Study 2 pre-registration can be accessed at https://osf.io/fex4t/?view_only=6aff1b71a75441e1bd1eec837c00219f.

Procedure

Up to 12 participants completed this in-lab study in each session, and a research assistant posed as a participant in sessions with an odd number of participants.

After consenting, participants were separated into individual cubicles and told that they would be communicate with another participant over the computer. As in Study 1, the study partner's responses were actually pre-programmed computer responses, and participants were not actually interacting with their study partner.

Participants first completed a short profile about themselves (e.g., name, major, hobbies), which was ostensibly sent to their study partner. In return, they received their study partner's supposed responses to the same profile.

Next, participants completed the stereotypic inference task (Czopp et al., 2006), which served as the context for the confrontation and as a pre-confrontation measure of stereotypic

responding. Described as a "photo-sentence" task, participants viewed a picture of a person and a brief description (e.g., "This person can be found in the theater"). Participants then generated a label to accompany that photo-sentence pair (e.g., "movie fan," or "actor"). Participants alternated with their study partner in providing the first response for each photo-sentence pair. Both the participant and the study partner provided a response to each pair. To make the pre-programmed responses more believable, participants saw "loading" symbols while their partner ostensibly completed their turn.

Embedded among 16 filler trials were three critical photo-sentence pairs that, because they were paired with images of Black men, could elicit stereotypic responses. For instance, a photo of a Black man paired with the sentence, "This person depends on money from the government," could elicit stereotypic responses, such as "homeless," or nonstereotypic responses, such as "civil servant." These critical pairs were strategically placed so that the participant always responded first to these items. Prior research (e.g., Burns, Monteith, & Parker, 2017; Chaney & Sanchez, 2018; Czopp et al., 2006) shows that participants typically respond with stereotypical answers, and that was indeed the case (M = 1.96, SD = .74).

Next, participants were randomly assigned to the *trust-present* (n = 154) or *trust-neutral* (n = 158) condition. Participants in the *trust-present* condition completed a "dilemma decision-making task" with their study partner. Specifically, participants wrote how they would respond to a particular dilemma. Their response was then sent to their study partner. Then, the study partner supposedly wrote a response for a second dilemma, which was sent to the participant. The study partner's response to the second dilemma showed that the study partner was willing to act in ways that benefit the participant, even though there was a cost to the study partner. For instance, one of the study partner's dilemmas read: "You are swiping through an online dating app when you see a profile of [Participant Name]'s significant other. You know [Participant Name] and their significant other are in a monogamous relationship. What do you do?" The study partner ostensibly responded: "Thats a tough one but I would tell [Participant Name] that I saw them on tinder or whatever the dating app was. It might be kinda awkward or uncomfortable, but its worth it if it means helping [Participant Name] find out what's going on." The study partner's response indicated willingness to do what is best for participant, even at the cost of discomfort. In total, there were six dilemmas: Three filler dilemmas that the participant responded to, and three critical

dilemmas that the study partner responded to, which were designed to increase the participant's trust in their study partner (see Appendix E).

In the *trust-neutral* condition, participants completed a getting-to-know-each-other filler task (Aron et al., 1997). Specifically, participants took turns responding to three open-ended questions (e.g., "For what in your life do you feel most grateful?") drawn from Aron et al.'s (1997) "Fast Friends" task. This getting-to-know-each-other filler task was an appropriate trust-neutral condition because participants interacted with their study partner as in the trust-present condition; however, that interaction was not designed to elicit trust. A pilot study using Cloud Research participants (n = 143; paid $2.25) and the same trust measure as in Study 1 indicated that the trust manipulation worked as intended. Specifically, participants in the trust-present condition trusted their study partner significantly more (M = 5.65, SD = .94) than participants in the trust-neutral condition (M = 4.85, SD = .89), $t(141) = 5.26$, $p < .001$, d = .87.

After completing the decision-making task, participants reported their trust of their study partner. We considered it possible that the trust manipulation would affect liking too, so participants also reported liking of their study partner so that we could control for its effects in analyses.

After reporting their liking and trust, participants were instructed to message their study partner with any reactions they had during the task. The study partner, who ostensibly also received these instructions, responded. Participants spent an average of 15.97 seconds (SD = 8.45) reading the partner feedback. Participants in the confrontation condition (n = 147) received a confrontation about using anti-Black stereotypes:

> i think the study went well but some of your answers about black people seemed kind of like stereotypes. like thinking about black people as criminals drug addicts (poor) and things like that. i just think sometimes its easy to jump to conclusions.

The confrontation was purposefully vague, so that it would be relevant to most participant responses, and differed slightly (indicated in parentheses) depending on which critical photo-sentence pairs the participant had received (see task details below).

Participants in the no-confrontation condition (n = 165) received neutral information about past studies:

> i think the study went well, but i've kind of done some other studies a little bit like this. Like a lot of these studies just have us sit on a computer, but it was cool we were paired up with someone for this one. I think it was good though.

Participants then completed measures of "how the interaction is going so far." Specifically, participants completed measures of affect, social costs, and nonprejudiced image threat.[3] After completing these measures, participants completed a second stereotypic inference task without their study partner. The second stereotypic inference task was described as a pilot test of new photo-sentence pairs for future studies and that researchers just wanted an idea of how people complete these new items. Participants were informed that they would not receive feedback on this task, and their responses would simply be added to an anonymous database. In reality, participant responses to this second stereotypic inference task served as a measure of stereotypic responding. Embedded within 17 filler trials were three new critical photo-sentence pairs for which stereotype-consistent responses were possible.

Finally, participants complete the demographic questions (e.g., race, gender), were probed for suspicion, and debriefed.

Measures

Affect, social costs, trust, and nonprejudiced image threat were assessed using the same items as Study 1, again using a 1 (*strongly disagree*) to 7 (*strongly agree*) scale (see Appendix F for analyses involving positive affect and discomfort).[4] Trust was assessed on a 1 (*strongly disagree*) to 9 (*strongly agree*) scale using the same six items as Study 1; however, given the trust manipulation, it was treated as a manipulation check, rather than a primary outcome measure. Unless otherwise noted, all measures will be completed on a 1 (*strongly disagree)* to 7 (*strongly agree*) scale.

[3] Trust was measured again after the confrontation. As in Study 1, the confrontation reduced trust ($M_{\text{confr}} = 4.40$, $SD_{\text{conf}} = 1.55$; $M_{\text{no-confr}} = 6.33$, $SD_{\text{no-onf}} = 1.03$), $F(307) = 193.28$, $p < .001$, $\eta_p^2 = .39$. However, this measure was not of theoretical interest and thus is not mentioned again.

[4] Desire to avoid contact was inadvertently omitted from the Study 2 survey and thus is not included in the social cost composite.

Liking. Participants completed four items created for the purposes of this study to assess the extent to which they liked their study partner (e.g., "I don't care for my study partner" [reverse-scored], "I like my study partner") on a 1 (strongly disagree) to 9 (strongly agree) scale. A principal components analysis using varimax rotation indicated that all items loaded onto 1 factor that explained 60.31% of variance (loadings > .68).

Stereotypic responding. There were two sets of critical photo-sentence pairs: Set A and Set B. The sets were counterbalanced, so that some participants received Set A before the confrontation and Set B after the confrontation, while other participants received Set B before the confrontation and Set A after the confrontation.

Set A contained the prompts, "This person can be found on the streets," "This person uses needles for recreation," and "This person can be found behind bars." Set B contained the prompts, "This person depends on money from the government," "This person is good at getting into locked doors, and "This person handles a lot of drugs."

Two researchers independently coded whether participant responses were stereotypic ($\kappa_{\text{pre-confrontation}} = .85$; $\kappa_{\text{post-confrontation}} = .95$). The researchers discussed any discrepancies and jointly made final decisions about stereotype use. We computed *pre-* and *post-stereotype percentages* by dividing the number of stereotyped responses generated by the number of possible stereotyped responses. Lower scores reflect less stereotypic responding.

Results

Analytic Procedure

Based on our pre-registered data analysis plan, we initially conducted a *t*-test to examine whether the trust manipulation influenced liking. Indeed, results revealed that participants in the trust-present condition liked their study partner more ($M = 7.97$, $SD = .89$) than participants in the trust-neutral condition ($M = 7.43$, $SD = 1.17$), $t(292) = 4.48$, $p < .001$, 95%CI[-.73, -.27], $d = .51$. In order to investigate the unique effects of trust, liking was included as a covariate in subsequent analyses.

Each dependent variable except stereotype use was analyzed using a 2 (confrontation: bias confrontation vs. no confrontation) x 2 (trust: trust-present vs. trust-neutral) ANCOVA. Following

past research (e.g., Burns & Monteith, 2019; Czopp et al., 2006), post-confrontation stereotypic responding was also analyzed using a 2 x 2 ANCOVA, but with pre-confrontation stereotype use entered as an additional covariate. Confidence intervals are for the mean differences between cells. Mediation and moderated mediation analyses were performed using Hayes' (2018) PROCESS (V3; Models 4 and 8, respectively). See Table 3 for reliability, descriptive statistics, and inter-measure correlations.

Manipulation check: Trust. We first considered the effect of the trust manipulation on self-reported trust. Results revealed a significant main effect of trust condition, $F(1, 307) = 44.81$, $p < .001$, $\eta_p^2 = .13$. Participants in the trust-present condition reported more trust ($M = 7.15$, $SD = .94$) in their study partner than participants in the trust-neutral condition ($M = 6.21$, $SD = 1.07$). Thus, the trust manipulation successfully boosted trust.

As expected, given trust was assessed before the confrontation manipulation, the main effect of confrontation, $F(1, 307) = 2.19$, $p = .14$, $\eta_p^2 = .007$, and the interaction between confrontation and trust, $F(1, 307) = .004$, $p = .95$, $\eta_p^2 < .001$, were not significant.

Negself. There was a significant main effect of confrontation, $F(1, 307) = 135.36$, $p < .001$, $\eta_p^2 = .31$. Participants who were confronted felt more negself ($M = 3.62$, $SD = 1.60$) than participants who were not confronted ($M = 1.86$, $SD = 1.05$). As expected, the main effect of trust condition, $F(1, 307) = .85$, $p = .36$, $\eta_p^2 = .003$, and the interaction between trust condition and confrontation, $F(1, 307) = 1.19$, $p = .28$, $\eta_p^2 = .004$, were not significant.

Negother. Results revealed a main effect of confrontation, $F(1, 307) = 225.74$, $p < .001$, $\eta_p^2 = .43$. Confronted participants felt more negother ($M = 3.25$, $SD = 1.53$) than participants who were not confronted ($M = 1.36$, $SD = .70$).

The main effect of trust condition on negother was not significant, $F(1, 307) = 2.73$, $p = .10$, $\eta_p^2 = .009$, and, contrary to hypotheses, the interaction between trust condition and confrontation on negother did not reach significance, $F(1, 307) = 3.10$, $p = .08$, $\eta_p^2 = .01$.

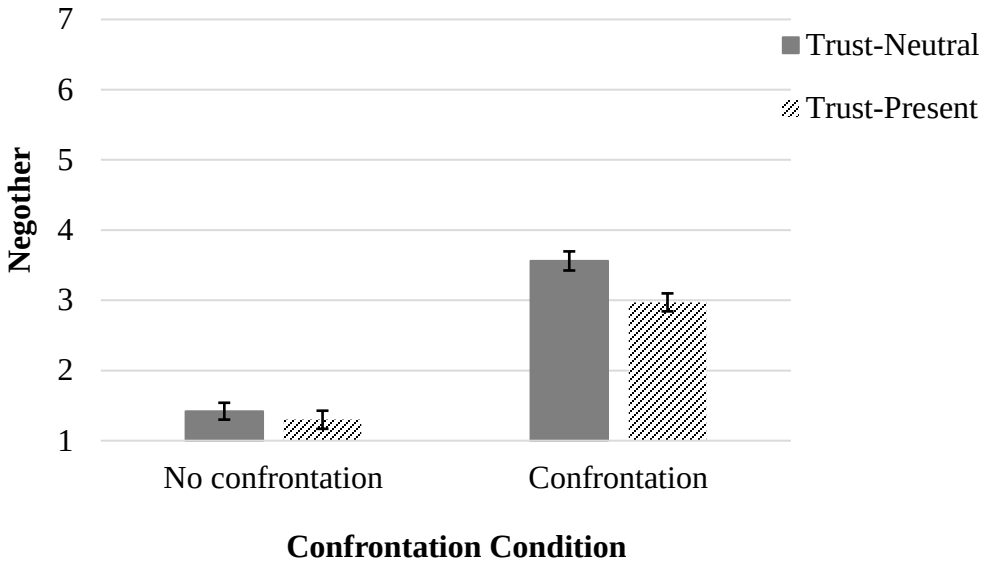
Table 3. Reliability, Descriptive Statistics, and Inter-Measure Correlations, Study 2

	α	*M (SD)*	1)	2)	3)	4)	5)	6)
Pre-Confrontation Measures								
1) Liking	.77	7.69 (1.07)	—	—	—	—	—	—
2) Trust	.78	6.67 (1.11)	0.63***	—	—	—	—	—
Post-Confrontation Measures								
3) NegSelf	.94	2.69 (1.60)	-0.02	-0.08	—	—	—	—
4) NegOther	.85	2.25 (1.50)	-0.19**	-0.13*	0.48***	—	—	—
5) Social Costs	.96	-.04 (0.96)	-0.23***	-0.23***	0.44***	0.78***	—	—
6) Image Threat	.87	3.80 (1.40)	-0.05	-0.10	0.53***	0.61***	0.73***	—
7) Stereotypic Responding	—	.43 (.37)	-0.006	0.02	-0.31***	-0.24***	-0.26***	-0.33***

Note. Liking and trust were completed before the confrontation; all other measures were completed after the confrontation. For correlations involving stereotypic responding, pre-confrontation stereotypic responding was controlled for, so partial correlations are reported. For social costs, the reliability is for the linear composite (Nunnally, 1978).

*$p < .05$. **$p < .01$. **$p < .001$.

However, given pre-registered hypotheses we examined the simple effects. Results were in line with hypotheses (see Figure 2). As expected, confronted participants who were in the trust-present condition reported significantly less negother than confronted participants who were in the trust-neutral condition, $t(307) = 2.38$, $se = .19$, $p = .02$, 95% CI[-.81, -.07], $d = .39$. Among participants who were not confronted, there was no difference between trust conditions, $t(307) = 1.03$, $se = .18$, $p = .97$, 95% CI[-.36, .34], $d = .17$. Thus, the confrontation increased feelings of anger and irritation towards the confronter, but trust in the confronter buffered against this effect.

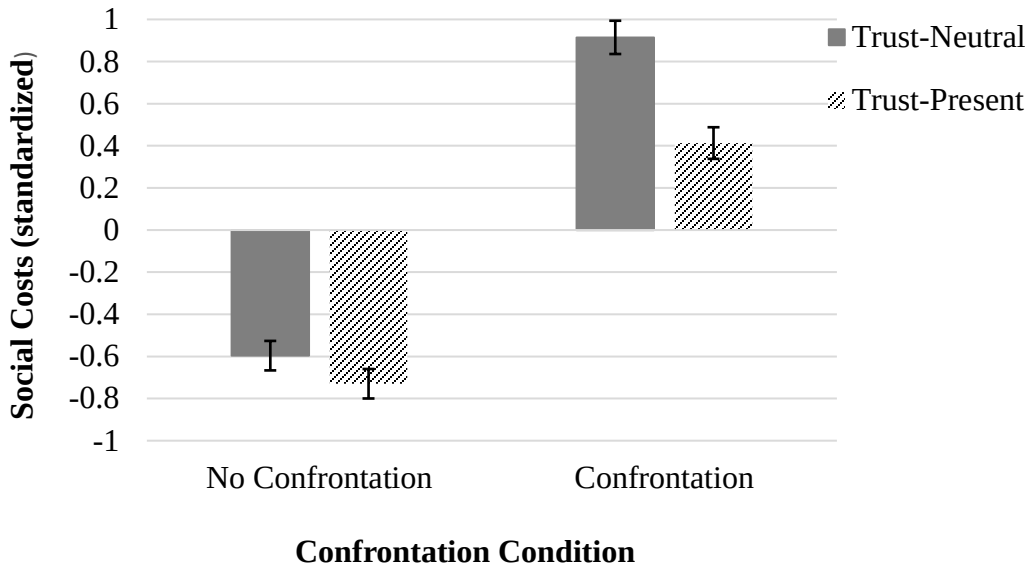


Note. Error bars represent +1/-1 standard error.

Figure 2. Negother as a function of confrontation and trust condition, Study 2.

Social costs. As with negother, results revealed a main effect of confrontation, such that confronted participants directed more social costs towards the confronter ($M = .65$, $SD = .92$) than not-confronted participants ($M = -.66$, $SD = .43$), $F(1, 307) = 321.24$, $p < .001$, $\eta_p^2 = .51$. A main effect of trust condition revealed that participants in the trust-present condition directed fewer social costs towards the confronter ($M = -.15$, $SD = .84$) than participants in the trust-neutral condition ($M = 0.06, 1.06$), $F(1, 307) = 7.63$, $p = .006$, $\eta_p^2 = .02$. More importantly, the interaction between trust condition and confrontation was significant, $F(1, 307 = 5.65$, $p = .02$, $\eta_p^2 = .02$ (see

Figure 3). The pattern of effects mirrored negother. As expected, confronted participants who were in the trust-present condition reported fewer social costs than confronted participants who were in the trust-neutral condition, t(307) = 3.41, se = .11, p < .001, 95% CI[-.60, -.17], d = .56. Among participants who were not confronted, there was no difference between trust conditions, t(307) = 2.01, se = .10, p = .74, 95% CI[-.17, .24], d =.32. In other words, the confrontation increased the extent to which the participant disliked and derogated the confronter; however, trust protected the confronter against this effect.

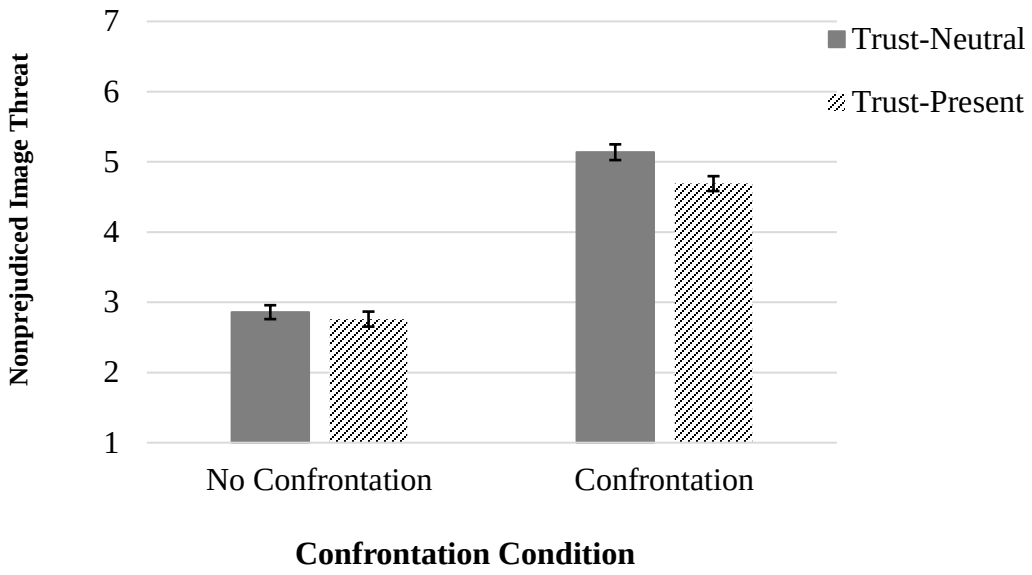


Note. Error bars represent +1/-1 standard error.

Figure 3. Social costs as a function of confrontation and trust condition, Study 2.

Nonprejudiced image threat. Results revealed main effects of confrontation, $F(1, 307) = 400.77$, $p < .001$, $\eta_p^2 = .57$, and trust condition, $F(1, 307) = 6.35$, $p = .01$, $\eta_p^2 = .02$. In line with Study 1, confronted participants reported more threat to their nonprejudiced image ($M = 4.90$, $SD = 1.12$) than not-confronted participants ($M = 2.82$, $SD = .74$). Furthermore, participants in the trust-present condition reported less image threat ($M = 3.72$, $SD = 1.34$) than participants in the trust-neutral condition ($M = 3.87$, $SD = 1.46$). The interaction between trust condition and confrontation was not significant, $F(1, 307) = 2.74$, $p = .10$, $\eta_p^2 = .009$. Nevertheless, simple

effects were in line with hypotheses (see Figure 4). As expected, confronted participants who were in the trust-present condition reported less image threat than confronted participants who were in the trust-neutral condition, t(307) = 2.65, se = .16, p = .004, 95% CI[-.75, -.14]. Among participants who were not confronted, there was no difference between trust conditions, t(307) = .43, se = .15, p = .50, 95% $CI_{difference}$ [-.19, .39]. Thus, the confrontation increased participant perceptions that the confronter threatened their nonprejudiced image; however, trust in the confronter buffered against this effect.



Note. Error bars represent +1/-1 standard error.

Figure 4. Nonprejudiced image threat as a function of confrontation and trust condition, Study 2.

Stereotypic responding. Thus far, Study 2 has demonstrated that trust alleviates confrontation-induced negother, social costs, and nonprejudiced image threat. These results indicate that confronters may wish to foster trust to avoid the interpersonal costs often incurred by confrontation. Yet, the ultimate goal of confrontation is to reduce bias. Was the confrontation successful at reducing bias?

As expected, no significant effects emerged in the analysis of pre-confrontation stereotypic responding, *F*s < 2.16, *p*s < .14. Thus, random assignment was effective. Pre-confrontation stereotypic responding explained a significant amount of variance when analyzing post-

confrontation stereotypic responding, $F(1, 306) = 12.30$, $p < .001$, $\eta_p^2 = .01$. More importantly, results revealed a significant effect of confrontation, $F(1, 306) = 81.96$, $p < .001$, $\eta_p^2 = .21$. In line with past research (e.g., Czopp et al., 2006; Chaney et al., 2018; Monteith et al., 2021), confronted participants used fewer stereotypes ($M = .25$, $SD = .31$) compared to participants who were not confronted ($M = .59$, $SD = .34$).

As expected, the main effect of trust condition was not significant, $F(1, 306) < .001$, $p = .99$, $\eta_p^2 < .001$. However, an unexpected interaction between trust and confrontation emerged, $F(1, 306) = 5.62$, $p = .02$, $\eta_p^2 = .02$. Confronted participants in the trust-neutral condition ($M = .30$, $SD = .33$) reduced their stereotypic responding more than not-confronted participants ($M = .55$, $SD = .35$), $t(306) = 4.57$, $se = .05$, $p < .001$, 95% CI[.14, .35], $d = .74$. Confronted participants in the trust-present condition ($M = .21$, $SD = .28$) also reduced their stereotypic responding more than not-confronted participants ($M = .64$, $SD = .33$), $t(306) = 8.72$, $se = .05$, $p < .001$, 95% CI[.32, .52], $d = 1.40$. However, this stereotype-reducing effect was larger among participants in the trust-present condition versus the trust-neutral condition. In other words, compared to participants who were not confronted, all confronted participants reduced their stereotypic responding; however, the bias-reducing effect of the confrontation was exaggerated among participants in the trust-present condition.

Relations Among Measures

The mediating role of negself on bias reduction. Given the nonsignificant interaction between confrontation and trust condition on negself, we conducted a simple mediation model in which confrontation was entered as a predictor, negself was entered as the mediator, post-confrontation stereotype use was entered as the outcome. Pre-confrontation stereotype use and liking were entered as covariates. Contrary to past research (Burns et al., 2017; Chaney & Sanchez, 2018; Czopp & Monteith, 2003; Czopp et al., 2006; Gulker et al., 2013; Parker et al., 2018), but in line with Study 1, the mediation model was not significant, $B = -.04$, $SE = .02$, 95%CI[-.08, .008].

The mediating role of negother and threat on social costs. Do negother and image threat explain the effect of trust and confrontation on social costs? Past research shows that negother mediates the relation between confrontation and social costs, such that confronted participants report more irritation and annoyance, which in turn leads to greater social costs. If trust alleviates social costs through negother, we would expect this mediational effect to be weaker in the trust-present versus the trust-neutral condition. We would expect the same pattern if image threat is a significant mediator.

To test these hypotheses, we conducted simultaneous moderated mediation analyses using Hayes's (2018) PROCESS (V3; Model 8; 5000 bootstraps). We entered confrontation and trust as the predictors, social costs as the outcome, and negother and image threat as the mediators. Although we did not pre-register negother due to an oversight, it is in line with Study 1 pre-registered analyses. The moderated mediation model for negother was not significant, B = -.13, SE = .08, 95%CI[-.29, .02]. This null effect is not surprising, given the null interaction (p = .08) between confrontation and trust on negother. Importantly, however, effects were in the expected direction: Specifically, the mediating effect of negother was weaker in the trust-present, B = .49, SE = .07, 95%CI[.35, .63], versus trust-neutral condition, B = .62, SE = .10, 95%CI[.44, .82]. This means that, among participants in both trust conditions, confrontation led to more negative other-directed affect than no-confrontation, which in turn led to greater social costs. This effect however was smaller, though not significantly, in the trust-present condition.

In line with Study 1, the moderated mediation model for image threat was also not significant, B = -.08, SE = .05, 95%CI[-.19, .02]. Like negother, this is likely due to the null interaction (p = .10) between confrontation and trust on image threat.

Discussion

Study 2 moves beyond correlational evidence to show that trust causes a reduction in confrontation-related social costs. Overall, confronted participants reported more social costs than not confronted participants; this effect, however, was attenuated among participants with greater trust in their study partner. These results are particularly exciting because they suggest that, in the real world, trust can be utilized to reduce social costs. For instance, armed with the knowledge that they will be protected from some of the social costs associated with the confrontation, would-be

confronters may feel more comfortable confronting trusted others. Alternatively, would-be confronters may proactively use this knowledge by fostering trust before a confrontation. Thus these results offer a clear, theoretically-supported strategy that would-be confronters can use to reduce social costs.

Why did trust reduce social costs? Results revealed mixed answers to this question. One the one hand, the moderated mediation models were not significant, and the interaction between confrontation and trust was not significant for both negative other-directed affect and image threat. On the other hand, for both measures, the interaction was trending, and all simple effects were significant in the hypothesized direction. In other words, despite the null interaction, confronted participants who trusted their study partner reported significantly less negative other-directed affect and image threat than confronted participants in the trust-neutral condition. Although the moderated mediation models were not significant, the simple mediation models for negative other-directed affect are consistent with Study 1. Specifically, among both trusted and trust-neutral dyads, confronted participants reported more negother, which in turn was associated with greater social costs. However, as expected, this pattern was weaker among the trust-present condition. Such findings are consistent with Study 1, which found that more trust was associated to less negative other-directed affect, which in turn was associated with fewer social costs.

Given Study 1 findings and the pattern of simple effects in Study 2, we are hesitant to rule out negative other-directed affect as an explanation for why trust reduces confrontation-induced social costs. Instead, we continued to explore the role of negative other-directed affect in Study 3. In contrast, given the nonsignificant mediation of image threat in both Studies 1 and 2, we can more confidently conclude that a reduction in image threat does not explain the relation between trust and confrontation-induced social costs. That is, even though trust reduces image threat, the reduction in image threat does not explain why trust reduces confrontation-related social costs. For these reasons, Study 3 did not consider nonprejudiced image threat and instead focused on the mediating role of negative other-directed affect.

STUDY 3

In Study 3, we investigated the relation between trust and confrontation-related social costs within an ecologically-valid context by using dyads who naturally differed in trust. Specifically, we examined the difference in social costs when participants believed they were confronted by a friend versus a stranger, and whether this difference was explained by the confrontee's trust of the confronter. We also continued to examine whether negative other-directed affect mediated the relation between trust and social costs.

Following past research, we hypothesized that, regardless of whether they have been confronted by a friend versus stranger, participants would evaluate their study partner more negatively and reduce subsequent biased responses after versus before the confrontation. More importantly, we hypothesized that the confrontee's relationship with the confronter (i.e., friends versus strangers) might moderate these effects. Although closeness makes people more sensitive to rejection, trust buffers against this effect (Murray et al., 2012). So, *compared to before the confrontation,* people may experience less negative other-directed affect and evaluate their partner less negatively after being confronted by a friend versus stranger. If this is the case, we further hypothesized that post-confrontation negative other-directed affect would mediate the effect of partner condition on post-confrontation social costs. Although we overlooked pregistering a mediational role for trust, it is clear from our rationale that we expected the effect of partner condition on social costs to be serially mediated by pre-confrontation trust and post-confrontation negative other-directed affect (i.e., partner condition → trust → negother → social costs).

We also considered an alternative, competing outcome. Specifically, we considered it possible that confrontations by close others may result in *greater* social costs than confrontations by strangers. Confrontations can be construed as a type of rejection, and rejection is more painful when it comes from close others than from strangers (e.g., Murray et al., 2006). Thus, compared to before the confrontation, people may feel more negative other-directed affect and report more social costs after being confronted by a friend versus stranger. If this is the case, we did not expect trust to mediate.

In sum, Study 3 advances our understanding of (a) how confrontations operate within an ecologically valid context (i.e., pre-existing relationships) and (b) whether trust reduces

confrontation-related social costs. The Study 3 pre-registration is available at https://osf.io/yzjak?view_only=a45d28e27ff84c5f997c6c88ebd47f23.

Method

Design and Participants

The study used a 2 (time: pre-confrontation vs. post-confrontation) x 2 (partner condition: friend vs. stranger) design. When determining effect size, we considered that (a) most effects in social/personality psychology range from small to medium (r = .22; Richard, Bond, & Stokes-Zoota, 2003), (b) confrontation research routinely produces medium-sized effects (d = .5; e.g., Czopp et al., 2006), and (c) simple effects are often smaller than omnibus effects (e.g., Giner-Sorella, 2018). Given these considerations, we conducted an a priori power analysis in GPower (Faul, Erdfelder, Buchner, & Lang, 2009) specifying a mixed ANOVA, a small effect size (f = .1, or d = .2,), a medium correlation among repeated measures (r = .30), and an alpha of .05. This analysis indicated that a minimum 278 participants (or approximately 50 participants per cell) would be needed for 80% power. However, given the planned mediation analysis and possible participant exclusions (e.g., suspicious participants; participants who don't believe they are actually interacting with their friend), we increased our target sample size to a minimum of 350 participants.

We recruited 500 non-Black undergraduate students (M_{age} = 18.86, SD_{age} = .96) for a study examining "the factors that influence group/dyad interaction."[5] Each participant received either partial course credit or $10 for participating.

Based on *a priori*, pre-registered decisions, data were excluded from six participants who did not give post-session consent to use their data (three in the friend condition, three in the stranger condition) and 57 participants who did not respond stereotypically prior to the confrontation (31

[5] Data collection began in January 2020 but was halted in March due to the COVID-19 pandemic. When data collection resumed in September 2020, the procedure was largely the same except (a) all participants wore masks and (b) were spaced six feet apart, including during face-to-face interactions. The intervening period between the two data collection periods (Spring 2020 and Fall 2020) also witnessed a resurgence of Black Lives Matter, a movement supporting racial justice (Wortham, 2020). A mixed analysis of variance with time (pre-confrontation vs. post-confrontation), partner condition (friend vs. stranger), and semester (Spring vs. Fall) entered as predictors revealed that the three-way interaction was not significant for all variables, Fs < 2.10, ps > .15.

in the friend condition, 26 in the stranger condition).[6] In addition, two researchers independently read participant responses to funnel debriefing questions and identified responses in which participants suspected they were not interacting with a real person. The researchers discussed these cases and jointly made final decisions about suspicion. Based on an *a priori*, pre-registered decision, 52 suspicious participants (29 in the friend condition and 23 in the stranger condition) were excluded from analyses. The final sample included 385 participants (73.2% women, 26.5% men, .3% gender-fluid; 75.1% White, 14.8% Asian, 4.4% Hispanic/Latinx).

Procedure

Participants were required to bring a same-gender friend (defined as someone they had known at least 2 months and had spent time with outside of class) to the research session to also participate in the study. Researchers verbally confirmed that the pair met this definition of friends at the beginning of the session. Sessions included between two and eight participants. Participants were randomly assigned to interact with either the friend they arrived with (the *friend condition)* or another participant (the *stranger condition*). When only two participants (i.e., one friend pair) signed up, those participants were automatically assigned to the friend condition.

Participants first completed a getting-to-know-each-other task with their study partner in a face-to-face interaction. Specifically, participants took turns responding to open-ended questions (e.g., "For what in your life do you feel most grateful?") drawn from Aron et al.'s (1997) "Fast Friends" task. However, participants completed only six of the 36 items from the "Fast Friends task," as our purpose was not to engender friendship but rather to provide some initial interaction. After five minutes, participants were moved into individual computer rooms, where they completed all measures (described below) aside from stereotypic responding. Afterwards, participants learned they would be interacting with their study partner again, but this time over the computer. As in Studies 1 and 2, the entire computer interaction was pre-programmed, and participants were not actually interacting with their study partner.

Then, participants completed the same stereotypic inference task as Study 2 (Czopp et al., 2006).

[6] Results did not significantly change when the responses of participants who did not respond stereotypically were included in analyses. Furthermore, excluded participants did not significantly differ from included participants on any demographic variable (i.e., gender, race, political orientation, internal motivation to respond without prejudice, and age).

After the photo-sentence task, participants were instructed to message their study partner with any reactions they had during the task. The study partner, who ostensibly also received these instructions, responded with the confrontation:

> I THINK THE STUDY WENT WELL BUT SOME OF YOUR ANSWERS ABOUT BLACK PEOPLE SEEMED KIND OF LIKE STEREOTYPES. LIKE THINKING ABOUT BLACK PEOPLE AS CRIMINALS DRUG ADDICTS (POOR) AND THINGS LIKE THAT. I JUST THINK SOMETIMES ITS EASY TO JUMP TO CONCLUSIONS

The confrontation was purposefully vague, so that it would be relevant to most participant responses, and differed slightly (indicated in parentheses) depending on which critical photo-sentence pairs the participant had received (see task details below). Participants spent an average of 17.40 seconds (*SD* = 8.33) reading the confrontation.

After the confrontation, participants completed all measures again, including a second stereotypic inference task. Participants also responded to demographic questions (e.g., race, gender) and questions about the friend they arrived with (i.e., "how long have you known the friend you came in with today?"). Finally, participants were probed for suspicion and debriefed.

Measures[7]

Unless otherwise noted, all measures were completed on a 1 (*not at all*) to 9 (*extremely/completely true*) scale. All measures had pre- and post-confrontation indices.

Affect, social costs, and trust were assessed using the same measures as Studies 1 and 2.[8] State attachment and closeness were also assessed but were not central to hypotheses (see

[7] At the end of the study, participants completed the five-item Internal Motivation to Respond Without Prejudice Scale (IMS; Plant & Devine, 1998; α = .80; e.g., "Because of my personal values, I believe that using stereotypes about Black people is wrong") as an exploratory measure on a 1 (*strongly disagree*) to 9 (*strongly agree*) scale. IMS did not moderate results and thus, consistent with our preregistration, it is not discussed further.

[8] As pre-registered, four additional items to assess unconditional positive regard (i.e., "I am confident that my study partner can look beyond my faults and see the best in me," "My study partner would like me to change some things about myself," "My study partner seemed irritated or impatient with some of my personal qualities," and "My study partner likes me unconditionally" were included in the study, with the thought that they might be combined with the trust items. However, confirmatory factor analyses suggested that trust and unconditional positive regard were two different factors, and an exploratory measurement invariance analysis revealed that strangers and friends respond differently to the unconditional positive regard items. For these reasons, we decided to exclude these four items from the trust index. Results did not change when these four items were included.

Appendix G for description and analyses). See Appendix H for analyses involving positive affect and discomfort.

Stereotypic responding. As in Study 1, the two sets of critical photo-sentence pairs were counterbalanced. Set A contained the same prompts as Study 1. Set B slightly differed: Due to a programming error, a problematic third prompt in Set B was included for the first 301 participants. This prompt was, "This person handles other people's money." This prompt was used in a prior confrontation study in our lab but was later discarded because it does not typically yield stereotypic responses. Indeed, it yielded stereotypic responses for only 2% of the participants who completed it in the present study. After the programming error was discovered, the item "This person handles a lot of drugs" was used as the third prompt for Set B.

As in Study 2, two researchers independently coded whether participant responses were stereotypic ($\kappa_{\text{pre-confrontation}} = .82$; $\kappa_{\text{post-confrontation}} = .78$). The researchers discussed any discrepancies and jointly made final decisions about stereotype use. We computed *pre-* and *post-stereotype percentages* by dividing the number of stereotyped responses generated by the number of possible stereotyped responses. This allowed us to exclude the problematic item "this person handles a lot of money;" in these cases, the divisor was two rather than three. Lower stereotype percentage scores reflect less stereotypic responding.

Results

Analytic Procedure

The pre-registered data analysis plan was to conduct one-way analyses of covariance (ANCOVA) with partner condition (0 = stranger, 1 = friend) entered as the between-participants variable and the relevant pre-confrontation measure entered as a covariate. However, we later saw the value of using mixed model analyses of variance (ANOVA) so that pre- and post-measures could be statistically compared. Thus, we report results based on 2 (time: pre-confrontation vs. post-confrontation) x 2 (partner condition: friend vs. stranger) mixed model ANOVAs. As pre-registered, we also performed multilevel analyses to account for nesting within pairs. However, results were redundant with the mixed model ANOVAs, and so only the mixed model ANOVAs

are presented below. Confidence intervals are for mean differences between cells. See Table 4 for descriptive statistics, reliability, and inter-measure correlations.

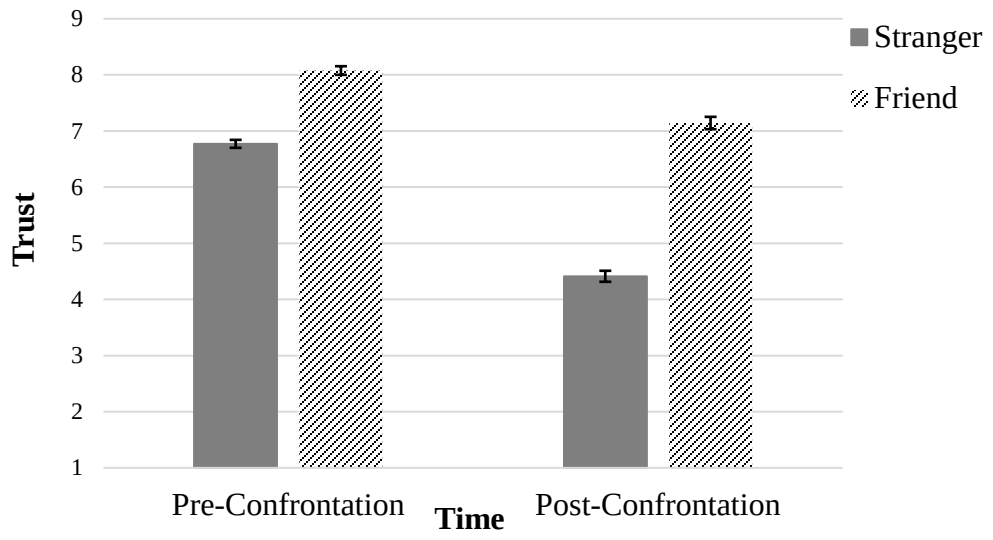
Trust. We first examined trust to establish that the friend and stranger conditions did indeed differ in their extent of pre-confrontation trust. Given we also assessed trust post-confrontation, we performed a mixed model ANOVA, as described above. A significant main effect for time was obtained, $F(1, 383) = 553.84$, $p < .001$, $\eta_p^2 = .59$, such that participants reported lower trust post-confrontation compared to pre-confrontation. A significant main effect for partner condition was also obtained, $F(1, 383) = 364.61$, $p < .001$, $\eta_p^2 = .49$, such that participants trusted their partner more in the friend than the stranger condition. In addition, the interaction was significant, $F(1, 383) = 103.56$, $p < .001$, $\eta_p^2 = .21$ (see Figure 5). Pre-confrontation, participants paired with a friend reported more trust than participants paired with a stranger, $t(383) = 16.24$, $p < .001$, 95%CI[1.10, 1.51], $d = 1.29$. This finding confirms that we tapped into differences in trust with the partner manipulation. Post-confrontation, strangers were also trusted less than friends, $t(383) = 51.05$, $p < .001$, 95%CI[2.44, 3.02], $d = 1.92$. However, the interaction emerged because the pre-to-post confrontation drop in trust was greater in the stranger condition, $t(383) = 58.58$, $p < .001$, 95%CI[2.17, 2.54], $d = 1.81$, than in the friend condition, $t(383) = 8.49$, $p < .001$, 95%CI[.73, 1.14], $d = .81$. This pattern suggests that confrontation reduces trust more when it occurs among strangers than among friends.

Table 4. Reliability, Descriptive Statistics, and Inter-Measure Correlations for Study 1

	α	*M(SD)*	1)	2)	3)	4)	5)	6)	7)	8)	9)
1) Pre NegSelf	.85	2.33 (1.33)	—	—	—	—	—	—	—	—	—
2) Post NegSelf	.93	4.26 (2.09)	0.29***	—	—	—	—	—	—	—	—
3) Pre NegOther	.76	1.74 (1.11)	0.64***	0.18**	—	—	—	—	—	—	—
4) Post NegOther	.81	3.02 (1.81)	0.36***	0.33***	.32***	—	—	—	—	—	—
5) Pre Social Costs	.89	-.40 (.47)	0.29***	0.05	.26***	.39***	—	—	—	—	—
6) Post Social Costs	.96	.44 (.98)	0.17**	0.22***	.15*	.67***	.61***		—	—	—
7) Pre Trust	.86	7.37 (1.21)	-0.20***	-0.07	-.11*	-.33***	-.67***	-.58***	—	—	—
8) Post Trust	.93	5.66 (1.97)	-0.10	-0.29***	-0.05	-.53****	-.47***	-.83***	.62***	—	—
9) Pre Stereotype %	—	74.76 (26.05)	-0.06	0.03	0.00	-0.07	-0.05	-0.05	.13**	0.06	—
10) Post Stereotype %	—	30.42 (31.67)	0.05	-0.16**	0.07	0.02	0.07	-0.01	0.04	0.06	.22***

Note. Measures labeled " Pre" refer to measures collected pre-confrontation. " Post" refers to measures collected post-confrontation. For social costs, the reliability is for the linear composite (Nunnally, 1978).

*$p < .05$. **$p < .01$. **$p < .001$.



Note. Error bars represent +1/-1 standard error.

Figure 5. Trust as a function of time and partner condition, Study 1.

Negself. Participants felt significantly more negself after the confrontation compared to before the confrontation, $F(1, 383) = 314.23$, $p < .001$, $\eta_p^2 = .45$. Furthermore, participants who were confronted by a stranger reported somewhat though not significantly more negself than participants who were confronted by a friend, $F(1, 383) = 3.49$, $p = .06$, $\eta_p^2 = .009$. Most importantly, the interaction between time and partner condition was significant, $F(1, 383) = 13.42$, $p < .001$, $\eta_p^2 = .03$, see Figure 6 (Panel A). As expected, the difference between the stranger and friend conditions was not significant at pre-confrontation, $t(383) = .93$, $p = .35$, 95%CI[-.14, .39], $d = .10$. Furthermore, negself increased from pre-confrontation to post-confrontation both for participants in the stranger condition, $t(383) = 15.81$, $p < .001$, 95%CI[-2.57, -2.00], $d = 1.28$, and the friend condition, $t(383) = 9.55$, $p < .001$, 95%CI[-1.82, -1.20], $d = .90$. This difference, however, was exaggerated for participants in the stranger condition, resulting in post-confrontation participants who were confronted by a stranger reporting significantly more negself than participants confronted by a friend, $t(383) = 3.10$, $p < .001$, 95%CI[-1.07, -.24], $d = .31$

Negother. Participants felt significantly more negother after the confrontation compared to before the confrontation, $F(1, 383) = 197.23$, $p < .001$, $\eta_p^2 = .34$. Furthermore, participants who were confronted by a stranger reported more negother than participants who were confronted by a friend, $F(1, 383) = 6.03$, $p = .01$, $\eta_p^2 = .02$. Most importantly, the interaction between time and partner condition was significant, $F(1, 383) = 28.46$, $p < .001$, $\eta_p^2 = .07$, see Figure 6 (Panel B). As expected, the difference between the stranger and friend conditions was not significant at pre-confrontation, $t(383) = 1.53$, $p = .13$, 95%CI[-.05, .40], $d = .15$. Furthermore, negother increased from pre-confrontation to post-confrontation both for participants in the stranger condition, $t(383) = 14.33$, $p < .001$, 95%CI[-1.96, -1.48], $d = 1.11$, and the friend condition, $t(383) = 5.91$, $p < .001$, 95%CI[-1.03, -.52], $d = 1.25$. This difference, however, was exaggerated for participants in the stranger condition, such that post-confrontation participants who were confronted by a stranger reported significantly more negother than participants confronted by a friend, $t(383) = 4.27$, $p < .001$, 95%CI[-1.13, -.42], $d = .44$. This post-confrontation difference in negother provides the first piece of evidence for the first hypothesis: Participants who were confronted by a friend, compared to those confronted by a stranger, felt less annoyed and irritated at the confronter post-confrontation.

Social costs. As with negother, results revealed a significant main effect of time, such that participants reported significantly more social costs towards the confronter after the confrontation compared to before the confrontation, $F(1, 383) = 506.23$, $p < .001$, $\eta_p^2 = .57$. Furthermore, as hypothesized, participants reported more social costs towards a stranger than a friend, $F(1, 383) = 167.76$, $p < .001$, $\eta_p^2 = .31$. Most importantly, the interaction between time and partner condition was significant, $F(1, 383) = 103.39$, $p < .001$, $\eta_p^2 = .21$ (see Figure 7). As expected, before the confrontation, participants paired with a stranger reported more social costs towards their study partner than participants paired with a friend, $t(383) = 8.24$, $p < .001$, 95%CI[-.46, -.28], $d = .85$. Furthermore, social costs increased after the confrontation, compared to before it, for both participants confronted by a stranger, $t(383) = 16.02$, $p < .001$, 95%CI[-1.27, -1.08], $d = 1.57$, and friend, $t(383) = 7.99$, $p < .001$, 95%CI[-.55, -.34], $d = .85$. However, the effect of time was exaggerated for participants in the stranger condition, resulting in greater post-confrontation social costs among participants in the stranger condition than in the friend condition, $t(383) = 13.12$, $p < .001$, 95%CI[-1.26, -.94], $d = 1.36$. In sum, although social costs increased after the

confrontation for all participants, friendship buffered this effect. It seems that friend confronters can escape a great deal of negative impressions and evaluations that arise as a result of confrontation.

Stereotypic responding. Results thus far show that, compared to participants confronted by strangers, participants confronted by friends reported more post-confrontation trust and less negother and social costs. These results are promising in terms of preserving interpersonal relationships after a confrontation. Yet, even if one wishes to preserve interpersonal relationships, the ultimate goal of confrontation is to reduce subsequent bias. Given participants confronted by friends also reported less negself than participants confronted by strangers, might the confrontation be less effective for stimulating subsequent bias regulation among friends than strangers?

Results indicated that, as expected, participants responded significantly less stereotypically post-confrontation compared to pre-confrontation, $F(1, 381) = 561.71$, $p < .001$, $\eta_p^2 = .60$. Furthermore, stereotypic responses did not vary by partner condition, $F(1, 383) = .17$, $p = .68$, $\eta_p^2 < .001$. Most importantly, the interaction was not significant, $F(1, 381) = .26$, $p = .61$, $\eta_p^2 = .001$. Thus, replicating past research, confrontation reduced stereotypic responding. Furthermore, these results show that, regardless of whether a friend or stranger performed the confrontation, the bias-reducing effect of confrontation remained strong.

Relation Among Measures

Negself and stereotypic responding. The pre-registered data analysis plan was to conduct a mediation analysis to determine whether negself mediates the relation between confrontation and stereotypic responding. However, we realized post hoc that such an analysis was not possible because all participants were confronted. Therefore, to test whether negative self-directed affect following confrontation was related to less biased responding, a regression analysis was performed in which post-confrontation stereotypic responding was predicted by post-confrontation negself, while controlling for both pre-confrontation stereotypic responding and pre-confrontation negself. As expected, we found a significant effect for post-confrontation negself, $B = -1.32$, $SE = .33$, $\beta = -.20$, $t(379) = 4.05$, $p < .001$, 95%CI[-1.97, -.68].

The mediating role of trust and negother on partner condition and social costs. Next, we conducted serial mediation analyses using Hayes' (2018) PROCESS (V3, Model 6, 5000 bootstraps) to test whether trust and negother mediate the effect of partner condition on social costs. Before conducting mediation analyses, we calculated difference scores for negother and social costs, which were then used in analyses (the results do not differ when a covariate rather than difference score approach is used.) Pre-confrontation trust was entered as a covariate in analyses.

Partner condition was entered as the predictor, pre-confrontation trust as the first mediator, negother difference score as the second mediator, and social costs difference score as the outcome. The serial mediation model was significant, $B = .07$, $SE = .03$, 95%CI[.02, .12]. As depicted in Figure 8, participants confronted by a friend, compared to those confronted by a stranger, reported more trust of the confronter, which was associated with less negother. Negother then predicted fewer social costs for the confronter. In other words, confrontees trusted friend confronters more than stranger confronters, which was associated with less annoyance and irritation about the confrontation. Those attenuated feelings of annoyance and irritation allowed the confronter to avoid some of the interpersonal costs typically associated with confrontation.

Discussion

Past research has focused almost exclusively on confrontations that occur between strangers; very little existing research has examined confrontations between friends (for an exception, see Brown et al., 2021). To our knowledge, no work has examined actual, versus imagined, friend-to-friend confrontations. Study 3 established that a confrontation by a friend reduced biased responding as effectively as a confrontation by a stranger, while simultaneously buffering against the social costs that confronters typically endure. Specifically, participants felt more negative-other directed affect and rated the confronter more negatively after the confrontation compared to before the confrontation. However, this effect was attenuated for participants who were confronted by a friend. Overall, people do not like to receive criticism, so confrontation still stings; however, when that confrontation comes from a trusted friend, that "sting" is not as strong. Mediation analyses suggest that, compared to stranger confronters, friend confronters are partially protected from social costs due to trust and an accompanying reduction

in negother: Participants confronted by a friend versus stranger reported more trust of the confronter, which was associated with less negative affect directed at the confronter, which in turn predicted fewer social costs.

Given the ultimate goal of confrontation is to reduce bias, we wish to emphasize that this reduction in social costs did not influence confrontation efficacy: The confrontation still reduced bias for participants in both partner conditions. Unexpectedly, people confronted by friends versus strangers did differ in post-confrontation negself: Compared to people confronted by strangers, people confronted by friends felt less guilt, self-criticism, and other forms of negative self-directed affect after the confrontation. Although not hypothesized, this post-confrontation difference for negself is consistent with the idea that we are not as sensitive to rejection from those we trust. As a result of this "softened" blow, people may experience less negself. Importantly, however, this difference in post-confrontation negself did not translate to a difference in stereotypic responding.

Would these results extend to other close others, such as parents, siblings, and romantic partners? We suspect so. Indeed, existing research shows that many constructs traditionally studied within romantic partners (e.g., self-disclosure, unconditional support, trust) operate similarly within friendships (Monsour, 1992). Along these same lines, constructs that operate within friendships should also operate within romantic partners and other close dyads. A caveat to this prediction is whether the close other has power over the confronter (e.g., a parent, a boss). Perceived social costs are greater when the confrontee has power over the confronter (Ashburn-Nardo et al., 2014); these heightened social costs may inhibit the moderating power of trust.

Overall, this research is important because friends may have more real-world opportunities to confront bias, given that people spend more time interacting with friends than with strangers. Furthermore, friends may also be more concerned than strangers that the confrontation will negatively impact their relationship. Study 3 somewhat alleviates these concerns by showing that friends who confront are actually more protected from social costs than strangers who confront, and that this protection did not change the bias-reducing effect of confrontation.

GENERAL DISCUSSION

The present research is the first to our knowledge to consider how interpersonal dynamics influence the confrontation context. Confrontations often occur in a dyad (i.e., one person confronting another person), and so many of the dynamics that apply to close relationships should also apply to the confrontation context. However, existing research has yet to consider confrontations from a close relationships perspective. By integrating these two previously separate research areas, the present research advances our understanding of the factors that influence confronter-confrontee relationships.

As a "first-step," correlational investigation, Study 1 showed that the more participants trusted their confronter, the less negative their impressions and evaluations of their partner (i.e., social costs); furthermore, negative other-directed affect mediated this effect. Study 2 then provided causal evidence that trust buffered confrontation's social costs. Finally, Study 3 showed that the effect of trust on social costs extends to an ecologically valid context: Confrontees reported fewer social costs in dyads with greater pre-existing trust (i.e., friends) than dyads with less pre-existing trust (i.e., strangers). Replicating Study 1, the effect of trust on social costs was again mediated by negative other-directed affect. Thus, across three studies and two types of bias (i.e., racism and sexism), the present research shows that interpersonal trust buffers against the social costs typically elicited by confrontations.

Why does trust buffer against social costs? We considered image threat as one such possibility. Specifically, we anticipated that trust would reduce the extent to which the confrontation elicits nonprejudiced image threat, which in turn would reduce social costs. Studies 1 and 2 revealed that trust reduces nonprejudiced image threat; however, this reduction in image threat did not mediate the relation between trust and social costs. Thus, the present research ruled out image threat as an explanatory mechanism. Perhaps, instead of a mediator, nonprejudiced image threat is better conceptualized as an additional form of social costs. Image threat and social costs were strongly correlated in both Study 1 ($r = .60$) and Study 2 ($r = .73$). Such high correlations suggest that image threat and social costs may be tapping into the same construct. Future research may explore this possibility via confirmatory factor analyses.

We considered negative other-directed affect as a second potential mediator. Past research indicates that negative other-directed affect mediates the relation between confrontation and social

costs (Monteith et al., in prep; Monteith et al., in press). For this reason, we anticipated that trust may reduce social costs by first reducing negative other-directed affect. Results supported this possibility. Specifically, in Studies 1 and 3, trust was associated with less negative other-directed affect, which in turn was associated with fewer confrontation-related social costs. Although the moderated mediation model was not significant in Study 2, negother mediated the relation between confrontation and social costs in both the trust-present and the trust-neutral condition. Taken together, these results suggest that trust reduces confrontation-related social costs by first reducing feelings of irritation and annoyance.

Practically, the present research offers a theoretically-grounded strategy that confronters can use to preserve positive impressions. Study 2 showed that trust can be fostered among strangers via a brief task, and Study 3 showed that confronters can draw upon pre-existing trust. This means that confronters who wish to mitigate social costs should either engage in trust-building behaviors or emphasize the trust that already exists between them and the confrontee.

The fluidity of trust makes it a particularly useful strategy. Past research has primarily focused on immutable factors that cannot be easily changed. For instance, past research shows that dominant-group confronters are targeted by lower social costs than marginalized-group confronters (Drury & Kaiser, 2014; Gulker et al., 2013; Rasinki & Czopp, 2010) and that social costs are lower when confronting racism versus sexism (Czopp & Monteith, 2003; Gulker et al., 2013). Yet in most cases, a would-be confronter cannot change their identity, nor can they change the type of bias that needs to be confronted. In contrast to these past findings, the present research offers a mutable factor that confronters can actively utilize to reduce social costs and preserve positive impressions.

Of course, fostering trust within a real-world context may be easier said than done. Particularly, statements and behaviors intended to induce trust may backfire if the confrontee feels the confronter is trying to manipulate them. In such cases, trust would not be successfully fostered and thus would not reduce the social costs associated with confrontation.

Time is another issue at play within real-world confrontations. Although, at five minutes, our manipulation of trust was relatively short, five minutes may still be more than one can spare within the confrontation context. Within the confrontation context, how can one briefly and succinctly foster trust? We consider establishing a common ingroup identity to be one such possibility. Existing research suggests a common ingroup identity (i.e., a superordinate identity

that includes both interaction partners, such that they share a common ingroup) increases trust between interaction partners (Andrighetto, 2012; Dovidio & Gaertner, 199; Riek et al., 2010). This can easily be incorporated into the confrontation context by reminding the confrontee of a shared group identity. Future research should examine this and other low-cost strategies for increasing trust.

Finally, future research should continue to examine confrontations from a dyadic, interpersonal perspective. Most existing confrontation research focuses on factors that solely influence either the confronter or the confrontee. For instance, is the confronter a target- or dominant-group member (e.g., Drury & Kaiser, 2014; Gulker et al., 2013; Rasinki & Czopp, 2010)? How strongly does the confrontee identify with their ingroup (e.g., Becker & Barreto, 2014; Kaiser et al., 2009)? These questions, though important, treat the confronter and confrontee as separate, non-interacting entities. In reality, the confrontee's words and actions influence the confronter, and vice versa. Thus, to understand the confrontation context fully, it is important to consider the dyadic, interactive nature of confronter-confrontee relationships. This dyadic perspective will allow researchers to examine new questions concerning the relationship between the confronter and confrontee. For example, how does the confrontee respond to the confrontation? How does this response influence the confronter? Just as the close relationships literature evolved from research on stable, individual differences in attraction to research on complex, interdependent relationships (Fletcher & Overall, 2010), perhaps too it is time for the confrontation literature to evolve from considering confronters and confrontees as mostly separate entities and instead to examine them from a dyadic, relational perspective. Drawing theoretical support from the close relationships literature, the present research represents an initial step towards this new, relational perspective of bias confrontations.

Beyond Trust to Other Interpersonal Factors

Beyond trust, what other interpersonal factors may improve confrontation outcomes? Anticipation of a shared future may be worthy of future study. In most existing confrontation research, the confrontation is "one-and-done;" that is, the confrontee is confronted by the confronter, and there is no subsequent interaction between the confronter and confrontee. However, anticipation of future interaction may cause the confrontee to reconstrue the confrontation so that there is less dissonance between the confrontee's feelings about the confrontation and the

knowledge that they will interact again. Although this factor has yet to be examined by existing research, it is in line the results of Mallett and Wagner (2011). This study is the only confrontation study of which we are aware to feature a post-confrontation interaction. Specifically, a female confederate confronted participants with either a gender-bias confrontation (i.e., "Are you assuming the nurse is female? That's kind of sexist don't you think?") or a gender-neutral confrontation (i.e., "I don't think that's such a good idea"). Participants then interacted with that confederate for a second time. Results revealed that, during the second interaction, participants who were confronted about gender bias compensated for the confrontation: They agreed more with the confronter, searched for common ground, and engaged in non-verbal behaviors that expressed liking (e.g., smiling). In other words, in contrast to past research demonstrating post-confrontation social costs, participants who were confronted about gender bias engaged in more positive interpersonal behavior than participants in the gender-neutral confrontation condition.

Why did participants in the gender-bias confrontation condition compensate more than participants in the gender-neutral confrontation condition? Anticipating a second interaction with the confronter may have created more dissonance among participants in the gender-bias confrontation condition than participants in the gender-neutral confrontation condition. Participants in the gender-bias confrontation condition may then have engaged in positive interpersonal behaviors to reduce that dissonance. In this way, anticipation of a future interaction may have led to more positive interpersonal behaviors. Although this possibility has yet to be empirically tested, it is in line with the idea that, when a shared future is anticipated, participants might direct fewer social costs towards the confronter. Overall, anticipation of a shared future is a ripe direction for future research.

Utilizing Trust to Reduce Bias

The present research focuses on how trust reduce social costs. Yet, it is also worth considering how trust influences bias reduction, or the extent to which one reduces their bias after a confrontation. After all, reducing subsequent expressions of bias is the ultimate goal of confrontation.

Trust may influence bias reduction by first influencing the perceived validity of the confrontation. Specifically, confrontees who trust the confronter may see the confrontation as more valid. In line with this possibility, past research indicates that, as trust in the feedback source

increases, people are more accepting of critical feedback (Earley, 1986) and alter their behavior more to be in line with that critical feedback (Dirks & Ferrins, 2001; Earley, 1988) In the same way, as trust in the confronter increases, confrontees may be more accepting of the criticism inherent in the confrontation and perceive the confrontation as more valid. According the Validity and Impugnment as Determinants of Other-Confrontation Consequences (VIDOCC) theory, perceived validity of the confrontation is an important antecedent of whether the confrontee reduces bias (Monteith et al., in press). By increasing perceived validity, trust may subsequently increase the extent to which the confrontation reduces bias.

The interaction between confrontation and trust on bias reduction observed in Study 2 is consistent with this possibility. Although all confronted participants reduced their bias, confronted participants in the trust-present condition reduced their bias more than confronted participants in the trust-neutral condition. However, we did not similarly find that a friend's confrontation reduced subsequent bias more than a stranger's confrontation, despite friends being trusted more than strangers. Additional research is needed to examine the effect of trust on the perceived validity of a confrontation, and whether there are downstream consequences for bias reduction.

Conclusion

Being disliked is anathema to the human condition (e.g., Baumeister & Leary, 1995). Yet, existing research suggests that being disliked is an inevitable consequence of bias confrontations. Thus, would-be confronters are often torn between two competing goals: They wish to confront biased behavior, but also to preserve a favorable relationship with the confrontee (Mallet & Melchiori, 2014). What, then, are would-be confronters to do? The close relationships perspective undertaken by the present work reveals a one answer: Trust. Specifically, fostering trust (or emphasizing pre-existing trust) allows confronters to preserve positive impressions while simultaneously reducing prejudice. Ultimately, trust is a powerful remedy to the barriers that might otherwise prevent confrontation and subsequent prejudice reduction.

REFERENCES

Andrighetto, L., Mari, S., Volpato, C., & Behluli, B. (2012). Reducing competitive victimhood in Kosovo: The role of extended contact and common ingroup identity. *Political Psychology, 33*(4), 513-529. https://doi.org/10.1111/j.1467-9221.2012.00887.x

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*(4), 596-612. https://doi.org/10.1037/0022-3514.63.4.596

Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., & Bator, R .J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin, 23*(4), 363-377. https://doi.org/10.1177/0146167297234003

Ashburn-Nardo, L., Blanchar, J. C., Peterson, 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

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*(3), 497-529. https://doi.org/10.1037/0033-2909.117.3.497

Becker, J. C., & Barreto, M. (2014). Ways to go: Men's and women's support for aggressive and nonaggressive confrontation of sexism as a function of gender identification. *Journal of Social Issues, 70*(4), 668-686. https://doi.org/10.1111/josi.12085

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

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

Bobo, L. D. (2001). Racial attitudes and relations at the close of the twentieth century. In N. J. Smelser, W. J. Wilson, & F. Mitchel (Eds.), *American becoming: Racial trends and their consequences* (pp. 264-301). Washington, DC: National Academy Press.

Brown, R. M., Craig, M. A., & Apfelbaum, E. P. (2021) European Americans' intentions to confront racial bias: Consider who, what (kind), and why. *Journal of Experimental Social Psychology, 95*, 104-123. https://doi.org/10.1016.j.jesp.2021.104123

Burns, M. D., & Granz, E. L. (2021). Confronting sexism: Promoting confrontation acceptance and reducing stereotyping through stereotype framing. *Sex Roles, 84*, 503-521. https://doi.org/10.1007/s11199-020-01183-5

Burns, M., & Monteith, M. J. (2019). Confronting stereotypic biases: Does internal versus external motivational framing matter? *Group Processes and Intergroup Relations, 22*, 930-946. https://doi.org/10.1177/1368430218798041

Burns, M. D., Monteith, M. J., & Parker, L. R. (2017). Training away bias: The differential effects of counterstereotype training and self-regulation on stereotype activation and application. *Journal of Experimental Social Psychology, 73*, 97-110. https://doi.org/10.1016/j.jesp.2017.06.003

Campbell, L., Simpson, J. A., Boldry, J. G., & Rubin, H. (2010). Trust, variability in relationship evaluations, and relationship processes. *Journal of Personality and Social Psychology, 99*(1), 14-31. https://doi.org/10.1037/a0019714

Carver, C. S., & Scheier, M. F. (1981). A control-systems approach to behavioral self-regulation. In L. Wheeler (Ed.), *Review of personality and social psychology, Vol. 2* (pp.107-140). Thousand Oaks, CA: Sage.

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., Alt, N. P., & Shih, M. J. (2021). The breadth of confrontations as a prejudice reduction strategy. *Social Psychological and Personality Science, 12*(3), 314-322. https://doi.org/10.1177/1948550620919318

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

Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially offensive behavior? *Psychological Science, 19,* 226-228. https://doi.org/10.1111/j.1467-9280.2008.02072.x

Cruwys, T., Greenaway, K. H., Ferris, L. J., Rathbone, J. A., Saeri, A. K., Williams, E., Parker, S. L., Chang, M. X-L., Croft, N., Bingley, W., & Grace, L. (2021). When trust goes wrong: A social identity model of risk taking. *Journal of Personality and Social Psychology, 120*(1), 57-83. https://doi.org/10.1037/pspi0000243

Czopp, A. M. (2013). The passive activist: Negative consequences of failing to confront anti-environmental statements. *Ecopsychology, 5*, 17-23.

Czopp, A. M. (2019). The consequences of confronting prejudice. In R. K. Mallett & M. J. Monteith (Eds.), *Confronting prejudice and discrimination: The science of changing minds and behaviors* (pp. 201-221). Cambridge, MA: 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

Devine, P. G., & Monteith, M. J. (1993). The role of discrepancy-associated affect in prejudice reduction. In D. M. Mackie & D. L. Hamilton (Eds.), *Affect, cognition, and stereotyping: Interactive processes in group perception* (p. 317-344). Cambridge, MA: Academic Press.

Deutsch, M. (1973). *The resolution of conflict: Constructive and destructive properties.* New Haven, CT: Yale University Press.

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

Dirks, K. T., & Ferrin, D. L. (2001). The role of trust in organizational settings. *Organization Science, 12*(4), 450-467. https://doi.org/10.1287/orsc.12.4.450.10640

Dovidio, J. F., & Gaertner, S. L. (1998). *On the nature of contemporary prejudice: The causes, consequences, and challenges of aversive racism.* In J. L. Eberhardt & S. T. Fiske (Eds.), *Confronting racism: The problem and the response* (pp. 3-32). Thousand Oaks, CA: Sage Publications, Inc.

Dovidio, J. F., & Gaertner, S. L. (1999). Reducing prejudice: Combating intergroup biases. *Current Directions in Psychological Science, 8*(4), 101-105. https://doi.org/10.1111/1467-8721.00024

Dovidio, J. F., & Gaertner, S. L. (2004*). Aversive racism.* In M. P. Zanna (Ed.), *Advances in experimental social psychology, Vol. 36* (pp. 1-52). Cambridge, MA: Elsevier Academic Press. https://doi.org/10.1016/S0065-2601(04)36001-6

Dovidio, J. F., Kawakami, K. L., & Gaertner, S. L. (2000). Reducing contemporary prejudice: Combatting bias at the individual and intergroup level. In S. Oskamp (Ed.), *Reducing prejudice and discrimination* (pp. 137-163). Mahwah, NJ: Erlbaum.

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

Earley, P. C. (1986). Trust, perceived importance of praise and criticism, and work performance: An examination of feedback in the United States and England. *Journal of Management, 12*(4), 457-473. https://doi.org/10.1177/014920638601200402

Earley, P. C. (1988). Computer-generated performance feedback in the magazine-subscription industry. *Organizational Behavior and Human Decision Processes, 41*(1), 50-64. https://doi.org/10.1016/0749-5978(88)90046-5

Eliezer, D., & Major, B. (2012). It's not your fault: The social costs of claiming discrimination on behalf of someone else. *Group Processes and Intergroup Relations, 15*(4), 487-502. https://doi.org/10.1177/1368430211432894

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

Fiske, S. T., & Stevens, L. E. (1993). What's so special about sex? Gender stereotyping and discrimination. In S. Oskamp & M. Costanzo (Eds.), *Claremont symposium on applied social psychology* (vol. 6, pp. 173-196). Thousand Oaks, CA: Sage Publications Inc.

Fletcher, G. J. O., & Overall, N. C. (2010). Intimate relationships. In R. F. Baumeister & E. J. Finkel (Eds.), *Advanced social psychology: The state of the science* (pp. 461-494). Oxford, United Kingdom: Oxford University Press.

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

Gaertner, S. L., & Dovidio, J. F. (1986). *The aversive form of racism.* In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (p. 61-89). Cambridge, MA: Academic Press.

Gardner, D. M., & Ryan, A. M. (2020). What's in it for you? Demographics and self-interest perceptions in diversity promotion. *Journal of Applied Psychology, 105*(9), 1062-1072. https://doi.org/10.1037/apl0000478

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

Gillath, O., Hart, J., Noftle, E. E., & Stockdale, G. D. (2009). Development and validation of state adult attachment measure (SAAM). *Journal of Research in Personality, 43*, 362-373. https://doi.org/10.1016/j.jrp.2008.12.009

Giner-Sorolla, R. (2018). Powering your interaction. *Approaching Significance.* https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/

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

Gulker, J. E., Mark, A. Y., & Monteith, M. J. (2013). Confronting prejudice: The who, what, and why of confrontation effectiveness. *Social Influence, 8*(4), 280-293. https://doi.org/10.1080/15534510.2012.736879

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis* (3rd ed.). New York, NY: 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,* 1318-1333. https://doi.org/10.1002/ejsp.2692

Kaiser, C. R., Hagiwara, N., Malahy, L. W., & Wilkins, C. L. (2009). Group identification moderates attitudes toward ingroup members who confront discrimination. *Journal of Experimental Social Psychology, 45,* 770-777. https://doi.org/10.1016/j.jesp.2009.04.027

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 and Intergroup Relations, 6*(3), 227-237. https://doi.org/10.1177/13684302030063001

Kawakami, K., Karmali, F., Vaccarino, E. (2019). Confronting intergroup bias: Predicted and actual responses to racism and sexism. In R. K. Mallett & M. J. Monteith (Eds.), *Confronting prejudice and discrimination: The science of changing minds and behaviors* (pp. 3-28). Cambridge, MA: Academic Press.

Kessler, R. C., Mickelson, K. D., & Williams, D. R. (1999). The prevalence, distribution, and mental health correlates of perceived discrimination in the United States. *Journal of Health and Social Behavior, 40*(3), 208-230.

Koudenburg, N., Kannegieter, A., Postmes, T., & Kashima, Y. (2020). The subtle spreading of sexist norms. *Group Processes & Intergroup Relations, 24*(8), 1467-1485. https://doi.org/10.1177/1368430220961838

Luchies, L. B., Wieselquist, J., Rusbult, C. E., Kumashiro, M., Eastwick, P. W., Coolsen, M. K., & Finkel, E. J. (2013). Trust and biased memory of transgressions in romantic relationships. *Journal of Personality and Social Psychology, 104*(4), 673-694. https://doi.org/10.1037/a0031054

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology, 67*(5), 808-817. https://doi.org/10.1037/0022-3514.67.5.808

Mae, L., & Carlston, D. E. (2005). Hoist on your own petard: When prejudiced remarks are recognized and backfire on speakers. *Journal of Experimental Social Psychology, 41*(3), 240-255. https://doi.org/10.1016/j.jesp.2004.06.011

Mallett, R. K., Ford, T. E., & Woodzicka, J. A. (2019). Ignoring sexism increases women's tolerance of sexual harassment. *Self and Identity, 20*(7), 913-929. https://doi.org/10.1080/15298868.2019.1678519

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

McMinn, M. R., Williams, P. E., & McMinn, L. C. (1994). Assessing recognition of sexist language: Development and use of the Gender-Specific Language Scale. *Sex Roles: A Journal of Research, 31*(11-12), 741-755. https://doi.org/10.1007/BF01544290

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology, 65*(3), 469-485. https://doi.org/10.1037/0022-3514.65.3.469

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

Monteith, M. J., Burns, M. D., & Hildebrand, L. K. (2019). Does the content of confrontations matter? What to say, how to say it, and associated outcomes. In R. K. Mallett & M. J. Monteith (Eds.), *Confronting prejudice and discrimination: The science of changing minds and behaviors* (pp. 225-248). Cambridge, MA: Academic Press.

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*, 267-288. https://doi.org/10.1207/s15324834basp1803_2

Monteith, M. J., Hildebrand, L. K., & Mallett, R. K. (in preparation). *Reduced bias and interpersonal costs: Two independent confrontation processes.*

Monteith, M. J., & Walters, G. L. (1998). Egalitarianism, moral obligation, and prejudice-related personal standards. *Personality and Social Psychology Bulletin, 24*(2), 186-199. https://doi.org/10.1177/0146167298242007

Monsour, M. (1992). Meaning of intimacy in cross- and same-sex friendships. *Journal of Social and Personal Relationships, 9*(2), 277-295. https://doi.org/10.1177/0265407592092007

Moser, C. E., & Branscombe, N. R. (2021). Male allies at work: Gender-equality supportive men reduce negative underrepresentation effects among women. *Social Psychological and Personality Science.* Online First Publication. https://doi.org/10.1177/19485506211033748

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

Murray, S. L., Bellavia, G. M., Rose, P., & Griffin, D. W. (2003). Once hurt, twice hurtful: How perceived regard regulates daily marital interactions. *Journal of Personality and Social Psychology, 84*(1), 126-147. https://doi.org/10.1037/0022-3514.84.1.126

Murray, S. L., Derrick, J. L., Leder, S., & Holmes, J. G. (2008). Balancing connectedness and self-protection goals in close relationships: A levels-of-processing perspective on risk regulation. *Journal of Personality and Social Psychology, 94*(5), 429-459. https://doi.org/10.1037/0022-3514.94.3.429

Murray, S. L., & Holmes, J. G. (2009). The architecture of interdependent minds: A motivation-management theory of mutual responsiveness. *Psychological Review, 116*(4), 908-928. https://doi.org/10.1037/a0017015

Murray, S. L., Holmes, J. G., & Collins, N. L. (2006). Optimizing assurance: The risk regulation system in relationships. *Psychological Bulletin, 132*(5), 641-666. https://doi.org/10.1037/0033-2909.132.5.641

Murray, S. L., Holmes, J. G., & Griffin, D. W. (2000). Self-esteem and the quest for felt security: How perceived regard regulates attachment processes. *Journal of Personality and Social Psychology, 78*(3), 478-498. https://doi.org/10.1037/0022-3514.78.3.478

Murray, S. L., Lupien, S. P., & Seery, M. D. (2012). Resilience in the face of romantic rejection: The automatic impulse to trust. *Journal of Experimental Social Psychology, 48*, 845-854. https://doi.org/10.1016/j.jesp.2012.02.016

Murray, S. L., Pinkus, R. T., Holmes, J. G., Harris, B., Gomillion, S., Aloni, M., Derrick, J. L., & Leder, S. (2011). Signaling when (and when not) to be cautious and self-protective; Impulsive and reflective trust in close relationships. *Journal of Personality and Social Psychology, 101*(3), 485-502. https://doi.org/10.1037/a0023233

Murray, S. L., Rose, P., Bellavia, G. M., Holmes, J. G., & Kusche, A. G. (2002). When rejection stings: How self-esteem constrains relationship-enhancement processes. *Journal of Personality and Social Psychology, 83*(3), 556-573. https://doi.org/10.1037/0022-3514.83.3.556

Nunnally, J.C. (1978). *Psychometric theory* (2nd edition). New York, NY: McGraw-Hill Publishing Company.

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

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

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

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

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*(1), 95-112. https://doi.org/10.1037/0022-3514.49.1.95

Rempel, J. K., Ross, M., & Holmes, J. G. (2001). Trust and communicated attributions in close relationships. *Journal of Personality and Social Psychology, 81*(1), 57-64. https://doi.org/10.1037/0022-3514.81.1.57

Richard, F. D., Bond, Jr., C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331-363. https://doi.org/10.1037/1089-2680.7.4.331

Riek, B. M., Mania, E. W., Gaertner, S. L., McDonald, S. A., & Lamoreaux, M. J. (2010). Does a common ingroup identity reduce intergroup threat. *Group Processes & Intergroup Relations, 13*(4), 403-423. https://doi.org/10.1177/1368430209346701

Righetti, F., & Finkenauer, C. (2011). If you are able to control yourself, I will trust you: The role of perceived self-control in interpersonal trust. *Journal of Personality and Social Psychology, 100*(5), 874-886. https://doi.org/10.1037/a0021827

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality, 35*(4), 651-665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

Rousseau, D. M., Sitkin, S., Burt, R. S., Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*(3), 393-404. https://doi.org/10.5465/amr.1998.926617

Rusbult, C. E., & Van Lange, P. A. M. (2008). Why we need interdependence theory. *Social and Personality Psychology Compass,* 2(5), 2049-2070. https://doi.org/10.1111/j.1751-9004.2008.00147.x

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

Schuman, H., Steeh, C., Bobo, L. D., & Krysan, M. (*1998). Racial attitudes in America: Trends and Interpretations.* Cambridge, MA: Harvard University Press.

Shallcross, S. L., & Simpson, J. A. (2012). Trust and responsiveness in strain-test situations: A dyadic perspective. *Journal of Personality and Social Psychology, 102*(5), 1031-1044. https://doi.org/10.1037/a0026829

Shelton, J. N., Richeson, J. A., Salvatore, J., & Hill, D. M. (2005). Silence is not golden: The intrapersonal consequences of not confronting prejudice. In *Stigma and group inequality: Social psychological perspectives* (pp. 65-81). London, England: Psychology Press. https://doi.org/10.4324/9781410617057

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

Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science, 16*(5), 264-268. https://doi.org/10.1111/j.1467-8721.2007.00517.x

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

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*(6), 845-851. https://doi.org/10.1037/0022-3514.89.6.845

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in Experimental Social Psychology, 21*, 261-302. https://doi.org/10.1016/S0065-2601(08)60229-4

Swim, J. K., Mallett, R., & Stangor, C. (2004). Understanding subtle sexism: Detection and use of sexist language. *Sex Roles: A Journal of Research, 51*(3-4), 117-128. https://doi.org/10.1023/B:SERS.0000037757.73192.06

Terry, D. J., Hogg, M. A., & White, K. M. (1999). The theory of planned behavior: Self-identity, social identity, and group norms. *British Journal of Social Psychology, 38*(3), 225-244. https://doi.org/10.1348/014466699164149

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. *Advances in Experimental Social Psychology*, *21*, 181-227. https://doi.org/10.1016/S0065-2601(08)60227-0

Thai, M., Lizzio-Wilson, M., & Selvanathan, H. P. (2021). Public perceptions of prejudice research: The double-edged sword faced by marginalized group researchers. *Journal of Experimental Social Psychology*, *96*, 104181. https://doi.org/10.1016/j.jesp.2021.104181

Trawalter, S., & Richeson, J. A. (2008). Let's talk about race, Baby! When Whites' and Blacks' interracial contact experiences diverge. *Journal of Experimental Social Psychology, 44*(4), 1214-1217. https://doi.org/10.1016/j.jesp.2008.03.013

U. S. Bureau of Labor Statistics. (2021, January 22). Labor force statistics from the current population survey. *United States Department of Labor*. https://www.bls.gov/cps/cpsaat11.htm

Vonasch, A. J., Reynolds, T., Winegard, B. M., & Baumesiter, R. F. (2018). Death before dishonor: Incurring costs to protect moral reputation. *Social Psychological and Personality Science, 9*(5), 604-613. https://doi.org/10.1177/1948550617720271

Vorauer, J. D., Main, K. J., & O'Connell, G. B. (1998). How do individuals expect to be viewed by members of lower status groups? Content and implications of meta-stereotypes. *Journal of Personality and Social Psychology, 75*(4), 917-937. https://doi.org/10.1037/0022-3514.75.4.917

Wang, K., Silverman, A., Gwinn, J.D., & Dovidio, J. F. (2015). Independent or ungrateful? Consequences of confronting patronizing help for people with disabilities. *Group Processes and Intergroup Relations, 18*(4), 489-503. https://doi.org/10.1177/1368430214550345

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

Woodzicka, J. A., Mallett, R. K., Hendricks, S., & Pruitt, A. V. (2015). It's just a (sexist) joke: Comparing reactions to sexist versus racist communications. *HUMOR: International Journal of Humor Research, 28*(2), 289-309. https://doi.org/10.1515/humor -2015-0025

Wortham, J. (2020, June 5). *A 'glorious poetic rage': This time is different. Here's why.* New York Times. https://www.nytimes.com/2020/06/05/sunday-review/black-lives-matter-protests-floyd.html

Wyer, N. A., Sherman, J. E., & Stroessner, S. J. (1998). The spontaneous suppression of racial stereotypes. *Social Cognition, 16*(3), 340-352. https://doi.org/10.1521/soco.1998.16.3.340

Wyer, N. A., Sherman, J. E., & Stroessner, S. J. (2000). The roles of motivation and ability in controlling the consequences of stereotype suppression. *Personality and Social Psychology Bulletin, 26*(1), 13-25. https://doi.org/10.1177/0146167200261002

Yoo, S. H., Clark, M. S., Lemay, Jr., E. P., Salovey, P., & Monin, J. K. (2011). Responding to partners' expression of anger: The role of communal motivation. *Personality and Social Psychology Bulletin, 37*(2), 229-241. https://doi.org/10.1177/0146167210394205

Zou, L. X., & Dickter, C. L. (2013). Perceptions of racial confrontation: The role of colorblindness and comment ambiguity. *Cultural Diversity and Ethnic Minority Psychology, 19*(1), 92-96. https://doi.org/10.1037/a0031115

APPENDIX A

Method

Participants were randomly assigned to the no-confrontation condition, the confrontation-only condition, or the image-protection confrontation condition. In the *image-protection confrontation condition*, participants read:

> I thought the task went okay, some of those dilemmas were tricky to respond to. But I noticed that for certain dilemmas you used "he" to refer to the person. Are you assuming the computer scientist, the CEO, the surgeon is a man? Women can have jobs like that too. *I'm sure you're NOT sexist or anything like that and that it was just a mistake.* I just think we just need to be careful not to make assumptions about gender.

The words in *italics* represent the portion of the confrontation that is specific to the image-protection condition. In the *confrontation-only condition*, participants received only the non-italicized words (i.e., "…But I noticed that for certain dilemmas you used "he" to refer to the person. Are you assuming the computer scientist, the CEO, the surgeon is a man? Women can have jobs like that too…"), without the italicized portion. In the *no-confrontation condition*, participants simply read the first sentence (i.e., "I thought the task went okay…").

Results

Analytic Procedure

Each dependent variable was be predicted using a one-way ANOVA (no-confrontation vs. confrontation-only vs. image-protection confrontation). All mediation and serial mediation analyses were conducted using Hayes' (2018) PROCESS (V3; Models 4 and 6 respectively; 5000 bootstraps); confrontation was recoded as necessary to make the appropriate comparison. See Table 5 for descriptive statistics as a function of condition.

Manipulation check. Results revealed a significant effect of confrontation, $F(2, 319) = 116.02$, $p < .001$, $\eta_p^2 = .42$. As expected, participants in the confrontation-only, $t(319) = 13.53$, $SE = .12$, $p < .001$, 95%CI[1.34, 1.80], $d = 1.82$, and image-protection conditions, $t(319) = 13.05$, $SE = .12$, $p < .001$, 95%CI[1.24, 1.71], $d = 1.77$, reported more nonprejudiced image threat than participants in the no confrontation condition. However, contrary to expectations, the difference

between the image-protection and confrontation-only conditions was not significant, $t(319) = .77$, $SE = .12$, $p = .43$, 95%CI[-.33, .14], $d = .11$. Thus the two confrontation conditions, compared to the no confrontation condition, successfully threatened the participant's nonprejudiced image; however, the image-protection manipulation, compared to the confrontation-only manipulation, did not successfully bolster the participant's nonprejudiced image. These results suggest that the reassuring the participant that they are not sexist is not enough to reduce nonprejudiced image threat.

Table 5. Measures as a Function of Confrontation Condition

	No Confrontation	Confrontation Only	Image-Protection Confrontation
Nonprejudiced Image Threat	2.90_a (0.80)	4.47_b (0.92)	4.38_b (0.87)
Negself	1.69_a (0.79)	2.59_b (1.34)	3.10_c (1.43)
Negother	1.32_a (0.75)	2.59_b (1.37)	2.73_b (1.34)
Social Costs	$-.60_a$ (0.57)	2.77_b (0.83)	0.34_b (0.77)
Trust	4.66_a (0.73)	3.63_b (0.85)	3.70_b (0.94)
Sexist Language Detection	5.64_a (4.66)	8.77_b (4.74)	9.14_b (5.17)

Note. Numbers in parentheses are cell standard deviations. Means not share a subscript with each dependent variable differ significantly at $p < .05$.

Negself. Once again, there was a significant effect of confrontation condition, $F(2, 319) = 38.99$, $p < .001$, $\eta_p^2 = .20$. As expected, participants in the confrontation-only, $t(319) = 6.17$, $SE = .16$, $p < .001$, 95%CI[.58, 1.22], $d = .82$, and image-protection conditions, $t(319) = 9.21$, $SE = .16$, $p < .001$, 95%CI[1.09, 1.74], $d = 1.23$, reported more negself than participants in the no confrontation condition. Surprisingly, given the failed image-protection manipulation, the confrontation-only and image-protection conditions also significantly differed, with participants in the image-protection condition reporting more negself than participants in the confrontation-only condition, $t(319) = 2.67$, $SE = .17$, $p = .002$, 95%CI[.19, .85], $d = .37$. Notably, however, this effect was much smaller than the effect of the confrontation conditions compared to the no-confrontation condition.

Negother. Results revealed a significant effect of confrontation condition, $F(2, 319) = 48.88$, $p < .001$, $\eta_p^2 = .24$. As before, participants in the confrontation-only, $t(319) = 8.64$, $SE = .16$,

$p < .001$, 95%CI[.96, 1.59], $d = 1.15$, and image-protection conditions, $t(319) = 9.72$, $SE = .16$, $p < .001$, 95%CI[1.10, 1.74], $d = 1.73$, reported more negother than participants in the no-confrontation condition. In other words, confronted participants felt much more annoyance and irritation at their study partner than non-confronted participants. Contrary to hypotheses, but in line with the failed image-protection manipulation, there was no difference between the confrontation-only and image-protection conditions, $t(319) = .75$, $SE = .16$, $p = .39$, 95%CI[-.18, .46], $d = .10$.

Social costs. There was again a significant effect of confrontation condition, $F(2, 319) = 58.22$, $p < .001$, $\eta_p^2 = .27$. The pattern of effects mirrored negother. Compared to participants in the no-confrontation condition, participants in the confrontation-only, $t(319) = 9.19$, $SE = .10$, $p < .001$, 95%CI[.68, 1.07], $d = 1.23$, and image-protection conditions, $t(319) = 10.33$, $SE = .10$, $p < .001$, 95%CI[.75, 1.14], $d = 1.39$, reported more social costs. Furthermore, as with negother, the confrontation-only and image-protection conditions reported the same amount of social costs, $t(319) = .58$, $SE = .10$, $p =. 53$, 95%CI[-.14, .26], $d = .08$. Thus, being confronted caused the participant to denigrate the confronter, regardless of whether the participant was assured that they weren't sexist.

Trust. Results revealed a now familiar pattern. There was a significant effect of confrontation condition, $F(2, 319) = 51.71$, $p < .001$, $\eta_p^2 = .25$, such that those in the confrontation-only, $t(319) = 9.25$, $SE = .11$, $p < .001$, 95%CI[.80, 1.24], $d = 1.30$, and image-protection conditions, $t(319) = 8.44$, $SE = .11$, $p < .001$, 95%CI[.73, 1.18], $d = 1.13$, reported less trust than those in the no-confrontation condition. As before, the confrontation-only and image-protection conditions did not significantly differ from one another, $t(319) = .54$, $SE = .11$, $p < .001$, 95%CI[.73, 1.18], $d = .08$. Thus, replicating Studies 1 and 2, being confronted reduced trust.

Sexist language detection. Once again, results revealed a significant effect of condition, $F(2, 319) = 17.20$ $p < .001$, $\eta_p^2 = .10$. As expected, both the confrontation-only, $t(319) = 4.95$, $SE = .66$, $p < .001$, 95%CI[1.83, 4.43], $d = .67$, and image-protection conditions, $t(319) = 5.20$, $SE = .66$, $p < .001$, 95%CI[2.19, 4.81], $d = .70$, increased sexist language detection compared to the no confrontation condition. Furthermore, as expected, sexist language detection did not differ between the two confrontation conditions, $t(319) = .53$, $SE = .68$, $p =. 59$, 95%CI[-.97, 1.71], $d = .07$. Thus, confrontation reduced biased responding compared to no confrontation.

APPENDIX B

Moral Decision-Making Task

Note: The participant responded to dilemmas 1, 3, and 5.

Dilemma 1: A CEO of a Fortune 500 company discovers that a long-time, high-ranking employee has been stealing from the company. What will the CEO need to do to respond to the situation?

Participant response was open-ended.

Dilemma 2: A waiter notices that a customer drops $100 from their pocket while leaving the table. The customer is new, and the waiter has to pay rent in a few days. What should the waiter do?

Study partner response: This is really tough lol. I know waiters aren't paid very much, and $100 is a lot of money. But it's also not right to keep something that's not yours. I think the waiter should return the money. Hopefully the customer will be thankful and give an extra tip, or become a regular customer.

Dilemma 3: *A computer programmer is instructed to use existing customer data to "train" a new software model. However, the computer programmer knows that doing so would violate the customer's right to privacy, according to the most recent Terms and Conditions. How should the computer programmer handle the situation?*

Participant response was open-ended.

Dilemma 4: A TA finds out that a student has cheated on a major exam. The student will fail the class without a high exam grade. How should the TA respond?

Study partner response: I feel bad for the student, I know what it's like to have so much riding on the exam. But the TA also has an obligation to report cheating. I think it depends how sure the TA is that the student is actually cheating, whether it is obvious or the TA just suspects. I guess just tell the professor about it and let the professor decide from there.

Dilemma 5: A surgeon finds out that a hospital patient has accidentally been given blood contaminated with the HIV virus. What should the surgeon do first?

Participant response was open-ended.

Dilemma 6: An employee discovers that the newly appointed department manager is the boss's cousin. The new manager doesn't have the appropriate qualifications and wasn't interviewed before being hired. How should the employee handle this situation?
Study partner response: I feel like nepotism is more common in the workplace than people like to admit. But, it's also not right if the new manager is bad at their job. I think the employee should report it to HR, HR probably knows how to handle it better than the employee.

APPENDIX C

Study 1 Measures

Note: The order of items within each scale were randomized. Items ending with (R) were reverse-scored.

Nonprejudiced Image Threat

Note: Critical items (indicated with **) will be embedded among filler items to distract from our true interest.

Instructions: Below are a number of traits that can describe a person. We are interested in **your thoughts on how your study partner thinks about you *at this moment.*** So, for each of the below questions, please respond based on **how you think your study partner feels about you *at this moment.***

At this moment, my study partner thinks of me as…

1.) Intelligent
2.) Independent
3.) Tolerant** (R)
4.) Sincere
5.) Confident
6.) Competitive
7.) Unbiased** (R)
8.) Daring
9.) Egotistical
10.) Agreeable
11.) Fair-minded** (R)
12.) Competent
13.) Friendly
14.) Dominant
15.) Skillful
16.) Prejudiced**
17.) Ambitious
18.) Nurturing

Affect (Monteith, 1993)

Instructions: Below are words that can describe different types of feelings. We are interested in your feelings <u>at this moment</u>. For each word, please indicate how much it describes your current feelings by selecting a number on the scale.

Not at all				Neutral				Extremely
1	2	3	4	5	6	7	8	9

At this moment, I feel….

1.) Disappointed with myself
2.) Guilty
3.) Embarrassed
4.) Self-critical
5.) Shameful
6.) Dissatisfied with myself
7.) Threatened
8.) Annoyed at others
9.) Irritated at others
10.) Happy
11.) Optimistic
12.) Fearful
13.) Friendly
14.) Good
15.) Proud
16.) Tense
17.) Uncomfortable

Social Costs

Partner Impressions (Czopp et al., 2006)

Instructions: For the next set of questions, *please consider your study partner. To what extent do you think each trait describes your study partner during the task you just completed?* Please remember that your study partner will not see your responses to this section, so your answers are completely anonymous.

Please be sure to respond based on how you feel about your study partner **at this moment.**

Not at all				Neutral				Extremely
1	2	3	4	5	6	7	8	9

1. Hypersensitive
2. A complainer
3. Hostile
4. Argumentative
5. Arrogant

6. Likeable (R)
7. Easy to work with (R)
8. Easy to get along with (R)
9. Nice (R)
10. Warm (R)
11. Intelligent*
12. Interesting*
13. Fast responder*

**indicates filler item.

Evaluation of the Interaction (Mallett & Wagner, 2011)

Instructions: Now you will answer several questions *about the interaction you just had moments ago with your study partner*. Please indicate the extent to which you agree or disagree with each statement using the scale provided below. Please remember that your study partner will not see your responses to this section, so your answers are completely anonymous.

Please be sure to respond based on how you feel about your study partner **at this moment.**

Not at all				Neutral				Extremely
1	2	3	4	5	6	7	8	9

1. I would be happy to work with my study partner on another task like the one we just dd in the future. (R)
2. I found it hard to get along with my study partner during this task.
3. I enjoyed working with my study partner very much during this task. (R)
4. My study partner was nice to me during this task. (R)
5. My study partner upset me during this task.
6. Thanks to my study partner, it was a pleasure to complete this task. (R)

Contact

Instructions: For the next set of questions, please consider *how much you want to interact with your study partner in the future.* Please remember that your study partner will not see your responses to this section, so your answers are completely anonymous.

Please be sure to respond based on how you feel about your study partner **at this moment.**

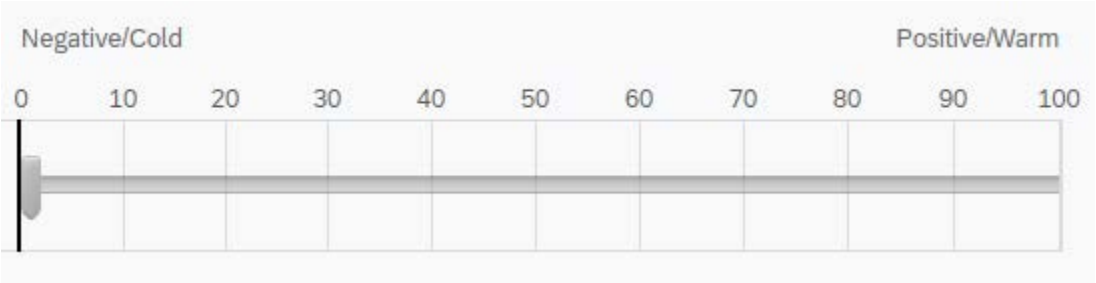
Not at all				Neutral				Extremely
1	2	3	4	5	6	7	8	9

1. Generally, how interested are you in making new friends (with people besides your study partner)?*
2. I want to hang out with my study partner outside this research setting. (R)
3. I don't want to interact with my study partner again.
4. I am interested in getting to know my study partner more than I already do. (R)

*only included in the pre-confrontation items, to use as a potential covariate.

Thermometer

Instructions: We can think of our feelings about others as falling on a thermometer, with 0 meaning our feelings are very cool and 100 meaning our feelings are very warm. *Please move the slider on the thermometer to a position that corresponds with your feelings about your study partner* **at this moment.** (R)



Trust

Instructions: For the next set of questions, *please consider how you feel about your study partner* **at this moment**. Please remember that the other participant will not see your responses to this section, so your answers are completely anonymous.

Please be sure to respond based on how you feel about your study partner **at this moment.**

Not at all true				Neutral				Completely true
1	2	3	4	5	6	7	8	9

1. I am confident that my study partner wants to work with me.
2. My study partner seems to have mixed feelings about me. (R)
3. My study partner seems to have doubts about whether or not we are compatible. (R)
4. My study partner seems willing to be supportive.
5. Even if I do something undesirable, my study partner seems willing to respond in a positive way.
6. Even if I do or say something flawed, I feel like I can rely on my study partner to see me in a positive way.

Nonprejudiced Image Threat

Note: Critical items (indicated with **) will be embedded among filler items to distract from our true interest.

Instructions: Below are a number of traits that can describe a person. We are interested in **your thoughts on how your study partner thinks about you *at this moment.*** So, for each of the below questions, please respond based on **how you think your study partner feels about you *at this moment.***

At this moment, my study partner thinks of me as…

1.) Intelligent
2.) Independent
3.) Tolerant** (R)
4.) Sincere
5.) Confident
6.) Competitive
7.) Unbiased** (R)
8.) Daring
9.) Egotistical
10.) Agreeable
11.) Fair-minded** (R)
12.) Competent
13.) Friendly
14.) Dominant
15.) Skillful
16.) Prejudiced**
17.) Ambitious
18.) Nurturing

Sexist Language Detection Task

Instructions: Read each of the following statements carefully and circle every writing problem you find, including problems with *grammar*, *spelling*, *punctuation*, *and sexist and otherwise discriminatory language*. Please work as quickly as you can. Try not to spend more than ten minutes working on the task. We want to know how many errors people can independently find in a short amount of time.

1. Twenty male and female participants were in a study on stress. Each persons' alertness was measured by the difference between his obtained relaxation score and his obtained arousal score.

2. The college basketball team was undefeated and ranked third in the naion, but the women's team had the worst record in the league.

3. The busness executive's learned about domestic tasks from the homemakers.

4. When making an important decision, one must first determine how other's will be affected and if the outcome is worth the cost.

5. The medical textbook noted that a surgical nurse must interact calmly with her patients in order to set them at ease priot to an operation.

6. The post office advertises that their mailmen aren't never late, no matter how bad the wether.

7. The men's room and the ladies' room were both closed for plumbing repears.

8. The supervisor talked individually with the employees who were to be laid off.

9. City planners know that conferences are often held in good vacaton locatons so conference attendees' wives and children can come as well.

10. The most recently hired secretary was asked to check her boses mail twice a day.

11. Evolutionary theory proposes that man is evolving thru a process of survival of the fittest.

12. When considering their own children, many parents wonder bout the effect that a childs position among his siblings has on his intellectual development.

13. The memorial was given in honer of the men who died while working in the coal mine when it callapsed.

14. The tall, black, person was speaking to the group.

15. The use of experiments in psychology presupposes the the mechanistic nature of man.

16. Mr. and Mrs. Charles Jones donated $100,000 to the University Library.

17. Students are required too write a total of four summary paper over the term.

18. Ten guys and seven girls went out for pizza after they completed thier fnal exams.

19. The brothers liked to play footbal in the evenings.

20. When each new client comes in, they are evalated as to their mental health.

21. The company rules indicate that the person elected chairman of the the board is to preside over board meetings.

22. A lawyer who is two much like his client can lose his objectivity.

23. Several employees was recognized for their eccellent service.

24. The printing company was looking for someone to work overtime by working saturday. They couldn't ask Jack, the most qualified man for the job, because he was away on vacation.

25. Many children need a lot of mothering in order to feel comfortable making transitionsfrom daycare to kindergarden.

26. The fire fighters' maintained composure when comfronted by the large dog.

27. Even thought they may have good intentions research scientists often neglect their wives and children.

28. The student asked, "How many questions are on the exam.

29. A communication theory indicates that, first, an individual becomes aroused buy violations of personal space, and than he attibutes the cause of this arousal to other people in his environment.

30. While nurses aids are frequently mistaken as nurses, male nurses aids are freuquently mistaken as Doctors.

APPENDIX D

Additional Study 1 Affect Analyses

Positive affect and discomfort were analyzed using separate independent t-tests (no-confrontation vs. confrontation). As expected, confronted participants reported less positive affect (M = 4.34, SD = .90) than participants who were not confronted (M = 5.03, SD = .80), $t(320) = 6.86$, $p < .001$, 95%CI [.49, .89], d = .80. Furthermore, confronted participants reported more discomfort (M = 2.57, SD = 1.27) than participants who were not confronted (M = 1.74, SD = 1.03), $t(281) = 6.40$, $p < .001$, 95%CI [-1.09, -.58], d = .70.

APPENDIX E

Study 2 Decision-Making Task

Note: Participant responded to Dilemmas #1, 3, and 5. Study partner responded to Dilemmas #2, 4, and 6.

Instructions: Now, you and Casey will complete another decision-making task together. Specifically, you and Casey will be presented with dilemmas that involve your study partner. So, you will respond to dilemmas involving Casey, and Casey will respond to dilemmas involving you. You will take turns providing a response to that dilemma. You have been randomly assigned to go first. You will write a response to the dilemma, and that response will be sent to Casey, who will read it. Then the process will repeat, but the roles will be switched: Casey will be presented with a dilemma and will provide a response, which you will read.

Dilemma #1: Casey is deciding between two jobs. One is a prestigious, high-salary position, but is far away from home. The other job doesn't pay as well, but is close to Casey's home and family. Casey has asked for your advice about which job to choose. What do you say?

Dilemma #2: You and [Participant's Name] are taking the same class. [Participant's Name] wants to pair up for a group project. However, you know that [Participant's Name] has been going through a tough time lately and might not be able to do their share of the work. So you might have to work a little bit extra if [Participant's Name] is your partner. What do you do?

Response: "That's a hard decision but i think i would probably give you the benfit of the doubt and partner up. I know how hard it is to keep up with your studies all the time and i could help out by doing extra work"

Dilemma #3: Casey's roommate recently left the room a complete mess, which broke one of the rules they had agreed on in their roommate contract. Casey is pretty upset, but torn on how to respond - the roommate had never done this before, but nevertheless broke a rule. What would you advise Casey to do?

Dilemma #4: You and [Participant's Name] are applying for the same job. The night before the application is due, you notice a big mistake in [Participant's Name]'s resume. You could tell [Participant's Name] about the mistake, which would hurt your chances of getting the job, or you could say nothing. What do you do?

Response: "i would definiteley tell you about the mistake, so that you have the chance to fix it before the interview starts. even if we're applying for the same job, id want to have your back and help you submit the best application possible"

Dilemma #5: Casey is looking to join a new extracurricular on campus. Casey must decide between a club sport or a social club. The club sport is more time consuming, but will help Casey stay healthy. The social club will help Casey make friends, but won't benefit Casey's physical health. Which club should Casey join, and why?

Dilemma #6: You are swiping through an online dating app when you see a profile of [Participant's Name]'s significant other. You know [Participant's Name] and their significant other are in a monogamous relationship. What do you do?

Response: "Thats a tough one but i would tell you that i saw them on tinder or whatever the dating app was. It might be kinda awkward or uncomfortable, but its worth it if it means helping you find out whats going on".

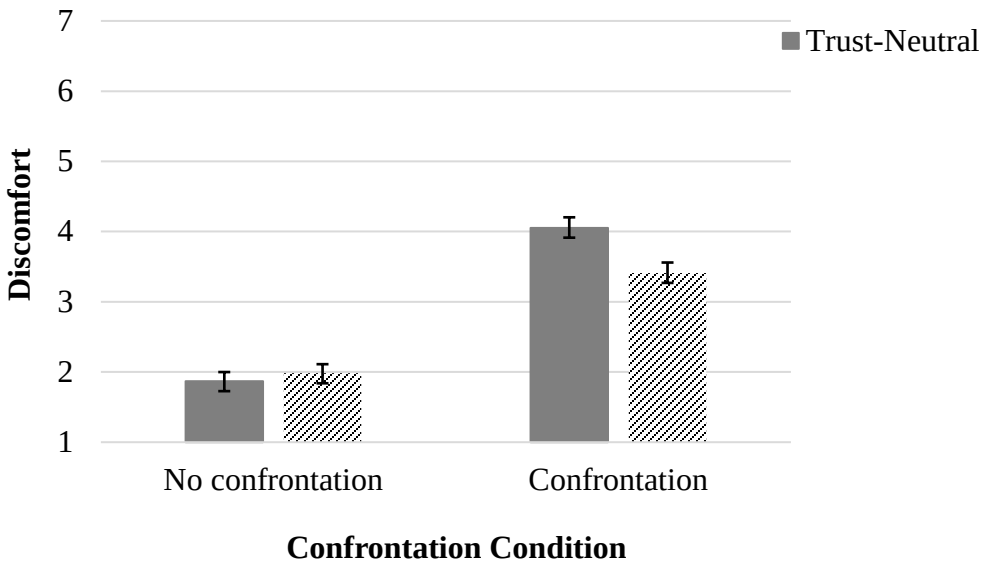
APPENDIX F

Additional Study 2 Affect Analyses

Positive affect (α = .88) and discomfort (α = .80) were analyzed using a 2 (confrontation: bias confrontation vs. no confrontation) x 2 (trust: trust-present vs. trust-neutral) ANCOVA, with liking included as a covariate.

Positive. There was a significant main effect of confrontation, $F(1, 307) = 127.76$, $p < .001$, $\eta_p^2 = .30$. Participants who were confronted felt less positive affect ($M = 3.88$, $SD = 1.08$) than participants who were not confronted ($M = 5.15$, $SD = .95$). The main effect of trust condition, $F(1, 307) = .50$, $p = .48$, $\eta_p^2 = .002$, and the interaction between trust condition and confrontation, $F(1, 307) = .30$, $p = .58$, $\eta_p^2 = .001$, were not significant.

Discomfort. Results revealed a main effect of confrontation, such that confronted participants felt more discomfort ($M = 3.71$, $SD = 1.44$) than not-confronted participants ($M = 1.92$, $SD = 1.06$), $F(1, 307) = 166.70$, $p < .001$, $\eta_p^2 = .35$. The main effect of trust condition was not significant, $F(1, 307) = 3.44$, $p = .06$, $\eta_p^2 = .01$. Most importantly, the interaction between trust condition and confrontation was significant, $F(1, 307) = 7.36$, $p = .007$, $\eta_p^2 = .02$ (see Figure 9). Confronted participants who were in the trust-present condition reported less discomfort than confronted participants who were in the trust-neutral condition, $t(307) = 2.87$, $se = 3.09$, $p = .002$, 95% CI[.24, 1.05], $d = .50$. Among participants who were not confronted, there was no difference between trust conditions, $t(307) = .57$, $se = .20$, $p = .57$, 95% CI[-.50, .27], $d = .06$. In other words, the confrontation increased the extent to which the participant felt discomfort; however, trust somewhat ameliorated this effect.



Note. Error bars represent +1/-1 standard error.

Figure 6. Discomfort as a function of confrontation and trust condition, Study 2.

APPENDIX G

State Attachment and Closeness

Measures

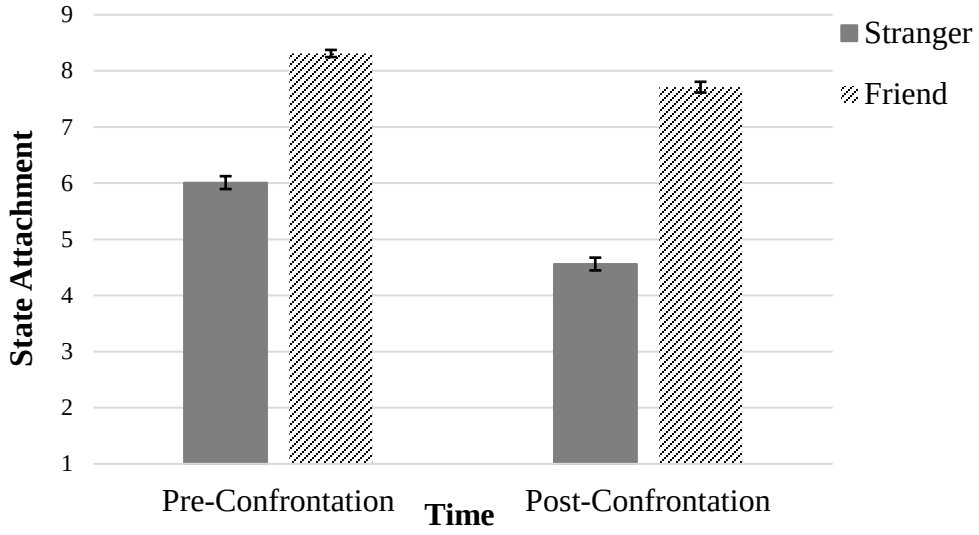
Participants completed three items adapted from the State Adult Attachment Measure (Gillath et al., 2009; e.g., "I feel like my study partner cares about me") to assess how secure and comfortable they feel towards their study partner. Participants also reported perceived closeness with their study partner using the single-item Inclusion of Other in the Self (Aron, Aron, & Smollan, 1992) pictorial scale.

Results

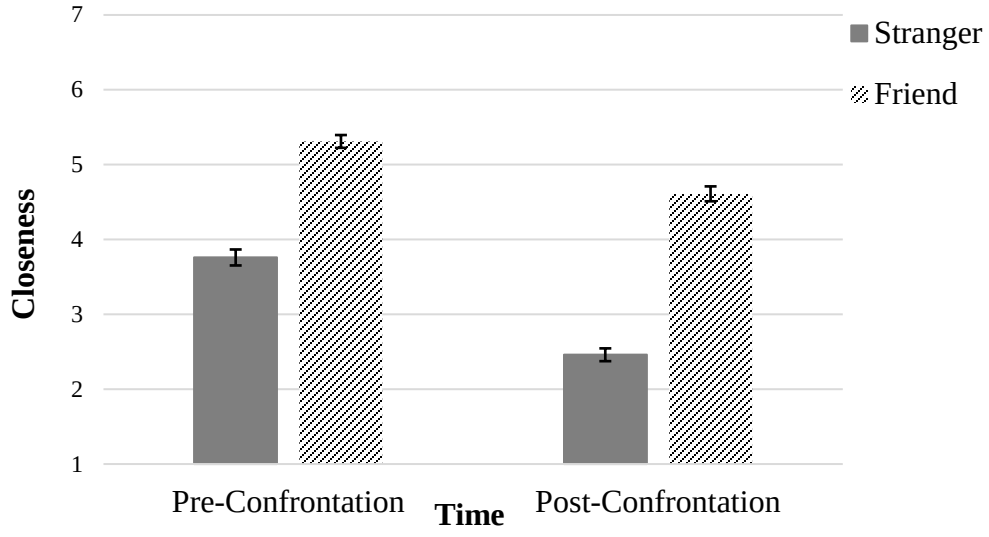
Each dependent variable was submitted to 2 (time: pre-confrontation vs. post-confrontation) x 2 (partner condition: friend vs. stranger) mixed model ANOVAs. Compared to before the confrontation, after the confrontation participants reported significantly less secure attachment, $F(1, 383) = 198.43$, $p < .001$, $\eta_p^2 = .34$, and closeness, $F(1, 383) = 243.41$, $p < .001$, $\eta_p^2 = .39$. Furthermore, participants paired with a friend reported significantly more secure attachment, $F(1, 383) = 472.77$, $p < .001$, $\eta_p^2 = .55$, and closeness, $F(1, 383) = 241.73$, $p < .001$, $\eta_p^2 = .39$, than participants paired with a stranger. More importantly, the interaction of time and partner condition was significant for both attachment, $F(1, 383) = 34.43$, $p < .001$, $\eta_p^2 = .08$, and closeness, $F(1, 383) = 21.84$, $p < .001$, $\eta_p^2 = .05$ (see Figure G1). As expected, before the confrontation participants paired with a friend reported more secure attachment, $t(383) = 16.66$, $SE = .14$, $p < .001$, 95%CI[2.03, 2.57], $d = 1.70$, and closeness, $t(383) = 11.10$, $p < .001$, 95%CI[1.28, 1.83], $d = 1.13$, than participants paired with a stranger. Furthermore, like trust, secure attachment and closeness dropped after the confrontation for both participants in the stranger (attachment: $t(383) = 14.76$, $p < .001$, 95%CI[1.26, 1.65], $d = .88$; closeness: $t(383) = 15.00$, $p < .001$, 95%CI[1.13, 1.47], $d = .93$) and friend (attachment: $t(383) = 5.57$, $p < .001$, 95%CI[.39, .81], $d = .56$; closeness: $t(383) = 7.42$, $p < .001$, 95%CI[.51, .88], $d = .57$) conditions. However, the drop was much more severe in the stranger condition than the friend condition. Specifically, after the confrontation participants paired with a stranger reported significantly less secure attachment, $t(383) = 20.79$, $p < .001$, 95%CI[2.86, 3.45], $d = 2.12$, and less closeness, $t(383) = 16.50$, $p < .001$, 95%CI[1.90, 2.41], $d =$

1.68, than participants paired with a friend. Thus, friendship with the confronter attenuated the negative interpersonal consequences of confrontation.

Panel A



Panel B



Note. Error bars represent +1/-1 standard error.

Figure 7. State attachment and closeness as a function of time and partner condition, Study 1.

APPENDIX H

Additional Study 3 Affect Analyses

Positive affect and discomfort were analyzed using separate 2 (time: pre-confrontation vs. post-confrontation) x 2 (partner condition: friend vs. stranger) mixed model ANOVAs. See Table 6 for descriptive statistics, reliability, and inter-measure correlations, and Table 7 for cell means and standard deviations.

Table 6. Reliability, Descriptive Statistics, and Inter-Measure Correlations, Study 1

	α	*M(SD)*	1)	2)	3)
1) Pre Positive	.82	6.85(1.16)			
2) Post Positive	.91	5.47(1.67)	.45**		
3) Pre Discomfort	.68	2.46(1.35)	-.33**	-.16*	
4) Post Discomfort	.80	3.64(1.94)	-.20**	-.61**	.43**

Notes. Measures labeled "Pre" refer to measures collected pre-confrontation. "Post" refers to measures collected post-confrontation.

*$p < .01$. **$p < .001$.

Table 7. Cell Means and Standard Deviations as a Function of Time and Partner Condition

	Friends		Strangers	
	Pre-Confr	Post-Confr	Pre-Confr	Post-Confr
Positive Affect	6.88_a (1.18)	5.99_b (1.52)	6.83_a (1.14)	5.19_c (1.66)
Discomfort	2.50_a (1.38)	2.43_b (1.34)	2.99_a (1.70)	4.19_c (1.96)

Note. Cell means are followed by standard deviation in parentheses. Means not sharing a subscript within each dependent variable differ significantly at $p < .05$.

Positive. Replicating past research (e.g., Czopp et al., 2006), participants felt significantly less positive after the confrontation compared to before the confrontation, $F(1, 383) = 308.31$, $p < .001$, $\eta_p^2 = .45$. Furthermore, participants who were confronted by a stranger reported less positive affect than participants who were confronted by a friend, $F(1, 383) = 16.65$, $p < .001$, $\eta_p^2 = .04$. Most importantly, the interaction between time and partner condition was significant, $F(1, 383) = 34.79$, $p < .001$, $\eta_p^2 = .08$. As expected, the difference between the stranger and friend conditions was not significant at pre-confrontation, $t(383) = .39$, $p = .70$, 95%CI [-.19, .28], $d = .04$. Furthermore, positive affect decreased from pre-confrontation to post-confrontation both for participants in the stranger condition, $t(383) = 17.35$, $p < .001$, 95%CI [1.58, 1.99], $d = 1.25$, and the friend condition, $t(383) = 7.92$, $p < .001$, 95%CI [.67, 1.11], $d = .65$). This difference, however, was exaggerated for participants in the stranger condition: Post-confrontation, participants who were confronted by a stranger reported significantly less positive affect than participants confronted by a friend, $t(383) = 5.75$, $p < .001$, 95%CI [.62, 1.27], $d = .59$. These results align with the results of both negself and negother: Participants who were confronted by a friend, compared to those confronted by a stranger, felt more positive post-confrontation.

Discomfort. Replicating past research (e.g., Czopp et al., 2006), participants felt significantly more discomfort after the confrontation compared to before the confrontation, $F(1, 383) = 163.15$, $p < .001$, $\eta_p^2 = .30$. Furthermore, participants who were confronted by a stranger reported more discomfort than participants who were confronted by a friend, $F(1, 383) = 16.10$, $p < .001$, $\eta_p^2 = .04$. Most importantly, the interaction between time and partner condition was significant, $F(1, 383) = 52.24$, $p < .001$, $\eta_p^2 = .12$. As expected, the difference between the stranger and friend conditions was not significant at pre-confrontation, $t(383) = .53$, $p = .60$, 95%CI [-.20, .35], $d = .05$. Furthermore, discomfort increased from pre-confrontation to post-confrontation both for participants in the stranger condition, $t(383) = 14.78$, $p < .001$, 95%CI [-2.00, -1.53], $d = 1.05$, and the friend condition, $t(383) = 3.77$, $p < .001$, 95%CI [-.74, -.23], $d = .32$). This difference, however, was exaggerated for participants in the stranger condition: Post-confrontation, participants who were confronted by a stranger reported significantly more discomfort than participants confronted by a friend, $t(383) = 6.35$, $p < .001$, 95%CI [-1.57, -.83], $d = .65$. These results align with the results of the other affect variables: Participants who were confronted by a friend, compared to those confronted by a stranger, felt less discomfort post-confrontation.