Journal of Experimental Social Psychology 88 (2020) 103832

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

A hero for the outgroup, a black sheep for the ingroup: Societal perceptions of those who confront discrimination

Maja Kutlaca*, Julia Becker[1], Helena Radke[2]

Faculty of Psychology, University of Osnabrück, Germany

ARTICLE INFO

Keywords:
Discrimination
Moral courage
Do-gooder derogation
Allies
Targets

ABSTRACT

Confrontation of discrimination can be seen as a form of morally courageous behavior, however those who engage in it presumably risk societal backlash. In three experiments, we examined societal perception of those who engage in confrontation of sexist (Study 1 and Study 2) and racist advertisement (Study 3). We tested two competing hypotheses. First, prior research on confrontation of discrimination suggests that members of disadvantaged groups who confront injustice (i.e., targets) should be judged more harshly than members of advantaged groups who confront (i.e., allies). Second, by drawing upon the insights from the work on do-gooder derogation, we proposed that allies and targets risk societal backlash, but more so from ingroup members than outgroup members. In Study 1, we found that disadvantaged group members evaluated an ally more positively than advantaged group members. Study 2 and Study 3 revealed that the audience had a positive view of confronters. However, members of advantaged groups supported allies less than targets, whereas members of disadvantaged groups preferred allies over targets (Study 2) or supported them equally (Study 3). Thus, our findings provide more support for the second hypothesis and we discuss their implications for the literature on moral courage, confrontation and do-gooder derogation.

In 2018, the highest award in American journalism, the Pulitzer Prize, was given to two newspapers: *The New York Times* for the work of Jodi Kantor and Megan Twohey, and *The New Yorker* for the work of Ronan Farrow. All three journalists drew public attention to harassment and sexual abuse in the film industry involving the producer Harvey Weinstein. In their articles, the journalists openly confronted sexism in the entertainment industry and not only gave a voice to Weinstein's accusers, but inspired others to come forward and speak openly about the abuse of power in politics, business, and science.

Confrontation of discriminatory treatment has important interpersonal and societal effects, because it can increase feelings of guilt among perpetrators and decrease future intentions to engage in discriminatory behavior (Czopp & Monteith, 2003; Czopp, Monteith, & Mark, 2006; Fazio & Hilden, 2001; Mallett & Wagner, 2011). At the same time, challenging others about their wrongdoings is costly (Dodd, Guiliano, Boutell, & Moran, 2001), and those who complain about injustice risk being punished by the perpetrator and labeled as 'complainers' by the broader society (Dodd et al., 2001; Kaiser, Hagiwara, Malahy, & Wilkins, 2009; Kaiser & Miller, 2001, 2003). Thus, confrontation of discriminatory treatment can be seen as a form of morally courageous behavior (Baumert, Halmburger, & Schmitt, 2013; Jonas & Brandstätter, 2004; Skitka, 2011), because it involves a reaction to a violation of social justice norms, and an attempt to influence and change the behavior of the perpetrator despite potential costs.

Majority of research on the confrontation of discrimination has focused on members of disadvantaged groups (i.e., *targets*), the likes of Jodi Kantor and Megan Twohey, and examined when and why they may lose or gain societal support for their actions (Becker & Barreto, 2014; Garcia, Schmitt, Branscombe, & Ellemers, 2010; Kaiser et al., 2009). Far less is known about whether members of advantaged groups like Ronan Farrow, who act on behalf of a disadvantaged group (i.e., as a male *ally*[3] to support women), can count on societal approval when

* Corresponding author at: Seminarstraße 20, 49074 Osnabrück, Germany.
E-mail address: maja.kutlaca@uni-osnabrueck.de (M. Kutlaca).

[1] This research has been funded through a grant awarded to Prof. Dr. Julia C. Becker on "Benefits and potential backlashes of allies engaging in solidarity-based collective action" by German Research Foundation (Deutsche Forschungsgemeinschaft): BE 4648/4-1.

[2] Helena Radke is now at Department of Psychology, The University of Edinburgh, UK.

We would like to thank our students and research assistants who helped with data collection. Furthermore, we are grateful to Prof. Linda Skitka and Maarten van Bezouw for their valuable feedback on this work.

[3] Allies in the confrontation literature are also referred to as non-target confronters.

https://doi.org/10.1016/j.jesp.2019.103832
Received 29 September 2018; Received in revised form 24 June 2019; Accepted 25 June 2019
Available online 12 July 2019

they confront discriminatory behavior (cf. Dickter, Kittel, & Gyurovski, 2012; Gervais & Hillard, 2014; Rasinski & Czopp, 2010). This paper takes a novel approach to understanding societal support for those who stand up against norm violations by integrating the insights from research on do-gooder derogation (Monin, 2007; Monin, Sawyer, & Marquez, 2008) with theory and research on the confrontation of discrimination. We propose that allies' and targets' confrontations may be met with societal (dis)approval, but this depends on the relationship between the confronters and their audience.

1. Confrontation of discriminatory treatment as an act of moral courage

The literature on moral courage distinguishes between a perpetrator who breaks a valued social norm, a victim, and a bystander who intervenes despite potentially negative consequences (Brandstätter, Jonas, Koletzko, & Fischer, 2016; Greitemeyer, Osswald, Fischer, & Frey, 2007; Jonas & Brandstätter, 2004). Importantly, what differentiates acts of moral courage from other pro-social behaviors is not the presence of risks per se, but the underlying motivation to restore a violated moral standard (Halmburger, Baumert, & Schmitt, 2015; Miller, 2000). In other words, to the extent that a bystander's confrontation of a discriminatory treatment is motivated by moral values (Kayser, Greitemeyer, Fischer, & Frey, 2010), it is considered to be morally courageous.

Despite strong norms against sexist or racist behavior in some societies, these behaviors are still present and even infiltrate public domains, such as advertising. For instance, in 2018 Swedish clothing company H&M released an advertisement featuring a young black male model wearing a sweatshirt that read “Coolest monkey in the jungle” (Bulman, 2018). Advertisements which use stereotypical depictions of disadvantaged groups have harmful consequences, because they can affirm and increase the perceptions of the targeted group as being less competent and not worthy of a moral treatment (Loughnan et al., 2010). Recognizing that such imagery transgresses valued moral and societal norms is an important way to reduce its negative impact.

The question remains whether those who recognize these norm violations and confront them are seen as heroes or self-righteous do-gooders by the broader society. According to the research on moral courage, those who intervene may face a variety of negative consequences ranging from verbal or physical retaliation from the perpetrator to loss of popularity, social distancing, derision and ostracism from the broader society (e.g., Bjørkelo, Einarsen, Nielsen, & Matthiesen, 2011; Sekerka & Bagozzi, 2007). Importantly, the literature on confrontation points out that the severity of the costs depends on the bystanders' group membership. There is evidence that allies' confrontations of sexist or racist behaviors are perceived less negatively by the perpetrators than targets' confrontations (e.g., Czopp & Monteith, 2003; Gulker, Mark, & Monteith, 2013). Less is known however, about societal reaction to targets and allies, and whether allies face less societal backlash than targets.

2. Reactions to confrontation

2.1. Perceptions of targets

For members of disadvantaged groups, expected reprisal is one of the key determinants of decisions not to confront discrimination (Ashburn-Nardo, Blanchar, Petersson, Morris, & Goodwin, 2014; Kaiser & Miller, 2001, 2003; Shelton & Stewart, 2004), and concerns over the societal reaction to confrontation decreases its likelihood (Good, Racusin & Sanchez, 2012; Kaiser & Miller, 2001; Shelton & Stewart, 2004). By engaging in confrontation, targets may confirm pre-existing negative stereotypes about their group (e.g., being difficult or aggressive), and risk being seen as violating societal norms about politeness (Dodd et al., 2001; Swim & Hyers, 1999). It is therefore not surprising that even though many members of disadvantaged groups anticipate that they would confront discrimination when it happens, few actually do (Hyers, 2007; Mallett & Melchiori, 2014; Shelton & Stewart, 2004; Swim & Hyers, 1999; Woodzicka & LaFrance, 2001).

2.2. Perceptions of allies

In contrast to target confronters who are particularly vulnerable to societal blowback, Drury and Kaiser (2014) proposed that allies may face less societal backlash than targets when they confront discrimination. We shortly review the evidence that speaks both in favor and against this hypothesis.

As mentioned earlier, several studies found that allies are not only judged less harshly by the perpetrators, they are also more effective in changing perpetrators' behavior (Czopp & Monteith, 2003; Gulker et al., 2013). Also, victims of discrimination (Cihangir, Barreto, & Ellemers, 2014), as well as the general audience (Eliezer & Major, 2012) tend to trust less bystanders who react to discrimination on behalf of someone else if the bystanders belong to the disadvantaged group than if they belong to the advantaged group. These studies however, only looked at bystanders who expressed their support to the victims, but did not confront the perpetrators. Moreover, Dickter et al. (2012) found that allies receive a lot of societal support for assertively confronting perpetrators, though they did not examine the support for target confronters in their studies. More direct evidence in support of the hypothesis comes from a study by Rasinski and Czopp (2010). The authors asked members of the advantaged group to watch a video featuring a White woman or a Black woman confronting a racist comment made by a White man. The audience evaluated the ally more positively than the target. But, the authors pointed out that less positive evaluations of the target confronter may have to an extent been driven by the “angry Black woman” stereotype (Landrine, 1985; Rasinski & Czopp, 2010). Such a negative stereotype does not apply to a Black male confronter, and the findings may not generalize to other domains of discrimination where negative stereotypes of target confronters may not exist (or at least not to the same extent).

However, there is also evidence that speaks against preferential evaluations of allies over targets. For instance, in another study examining the impact of interpersonal confrontation in the lab, Czopp et al. (2006) did not find any difference in the effectiveness between a Black and a White confronter in terms of decreasing stereotypical responses among the perpetrators. Surprisingly, a Black confronter was found to be more successful than a White confronter in eliciting self-directed negative feelings among the perpetrators (for more details see Study 2, Czopp et al., 2006). Moreover, a study examining heterosexuals' (i.e., general audience) support for gay and straight confronters of anti-gay prejudice yielded no differences between the evaluations of targets and allies (Cadieux & Chasteen, 2015). Lastly, Gervais and Hillard (2014) examined societal perceptions of female and male leaders who confronted a sexist remark either publicly or privately, and using a direct (i.e., labelling the comments as sexist) or an indirect style of confrontation (i.e., labelling the comments as unfair). Importantly, the results revealed that allies had an advantage over targets, but only when they confronted indirectly in a public context. In contrast, using a direct confrontational style in public, or confronting perpetrators privately (either directly or indirectly), did not make allies more popular than targets.

Thus, there is evidence (albeit limited and inconclusive) for the hypothesis that allies may receive greater societal support than targets. Our research contributes to the literature first by directly comparing how allies and targets are perceived by the general audience in two different contexts (i.e., sexism and racism). Moreover, we go beyond previous theory and research to propose that positive evaluations of allies and targets depend on the match between confronters and their audience.

3. Group dependent evaluations of targets and allies

After observing those who engage in moral acts, individuals often report feelings of generosity, heroism, and intention to help others themselves (Algoe & Haidt, 2009; Haidt, 2003; Schnall, Roper, & Fessler, 2010). However, admiration can sometimes be tainted by feelings of resentment. For instance, research on do-gooder derogation shows how moral exemplars (i.e., people committed to moral ideals or principles; Walker, 1999), can threaten others' moral identities and sense of moral self-worth that in turn leads to disapproval and derogation (Cramwinckel, van Dijk, Scheepers, & van den Bos, 2013; Minson & Monin, 2012; Monin et al., 2008; Monin & Jordan, 2009). Monin et al. (2008) found that people who went along with a racist task derogated an individual who refused to engage in it. Derogation occurs because people imagine that moral exemplars see themselves as morally superior and others as morally inferior (Minson & Monin, 2012), and/or anticipate that they will be reproached by the morally superior individual (O'Connor & Monin, 2016).

In his theoretical work, Monin (2007) proposed that individuals should experience a stronger threat to their moral identities when they compare themselves to superior others who are more (rather than less) similar to them (cf. Major, Testa, & Bylsma, 1991; Mussweiler, 2003; Wood, 1989). In intergroup settings, such as the confrontation of discrimination, members of the same group usually perceive greater similarity among each other (Brewer, 1979). This implies that for members of advantaged groups, allies' actions should be a more relevant standard of comparison and therefore more threatening than targets' actions, whereas the opposite should be true for disadvantaged groups. Consequently, we propose that allies and targets will be more likely devalued for their moral behavior by ingroup rather than by outgroup members (*Hypothesis 2*).

There are a few additional hints in the literature suggesting the plausibility of this hypothesis. For instance, research on the 'black sheep effect' (Marques & Paez, 1994) suggests that while ingroup members are generally liked more than outgroup members, this does not apply to those ingroup members who are perceived to be antinormative (Abrams, Marques, Bown, & Henson, 2000). Similarly, research on ingroup and outgroup criticism (i.e., criticism directed at changing the group's norms or behavior) suggests that ingroup critics are more likely to be taken seriously than outgroup critics (i.e., Intergroup Sensitivity Effect; Hornsey & Imani, 2004; Hornsey & Esposo, 2009). However, if the ingroup is criticized publicly (Elder, Sutton, & Douglas, 2005), or the intergroup conflict is made salient (Ariyanto, Hornsey, & Gallois, 2010), the preference for ingroup critics disappears. Thus, when the actions of ingroup members are seen as threatening to group norms (Abrams et al., 2000), or to group unity (Hornsey & Esposo, 2009), those ingroup members may be derogated as much or perhaps even more than critical outgroup members. This fits with the findings on do-gooder derogation. However, research on do-gooder derogation conceptualizes the negative response as a result of a threat to individual moral identities rather than to cherished group identities. We return to this point in the General Discussion.

All in all, the insights from the literature on do-gooder derogation and ingroup deviance further challenge the idea that allies may be unequivocally liked more than targets. Rather, this line of research points to the possibility that allies and targets may both inspire, but also threaten others.

4. Overview and hypotheses

We conducted three studies to address the question of the general audience's evaluation of those who act to restore violated social norms. We examined the perceptions of allies (Study 1), as well as allies and targets (Study 2) who spoke against derogatory advertisement targeting women in Germany, or against derogatory advertisement targeting Black Americans in the United States (Study 3). The goal of Study 1 was to explore societal views of male allies who confront sexism, and to identify key factors that lead to positive and/or negative perceptions of allies. In Study 2 and Study 3, we compared perceptions of targets and allies and tested two competing hypotheses. In line with the research on confrontation of discrimination (Drury & Kaiser, 2014), *Hypothesis 1* posits that, across the board, allies' confrontations will be met with less societal backlash than targets' confrontations. Alternatively, based on the insights from the literature on do-gooder derogation (Monin, 2007), *Hypothesis 2* qualifies this prediction and suggests that allies and targets risk societal backlash, but more so from their own fellow group members than from outgroup members.

5. Study 1

Prior work on societal perceptions of men who confront sexism is rather rare (cf., Gervais & Hillard, 2014). In Study 1, we approached the topic in an exploratory manner with the aim to clarify some opposing findings in the literature and develop our research design. First, although Gervais and Hillard (2014) found that male confronters were evaluated positively, other research showed that men who are feminists were perceived as weak and too feminine (Rudman, Mescher, & Moss-Racusin, 2013). The reason why a man who confronts sexism may be liked is because he endorses gender equality beliefs, and not necessarily because he acts upon them. Gervais and Hillard (2014) did not include a control condition in their study, hence it is not known whether the ally was liked for his actions or his beliefs. In order to do so, we compared a male ally confronter to two non-confronters: a) a non-confronter who endorses gender equality and perceives a norm violation (e.g., sees the advertisements as offensive to women), and b) a non-confronter who does not perceive a norm violation nor acts.

Second, prior research suggests that the style of confrontation adopted by the ally has an impact on societal perceptions. While some studies find that a more indirect style of confrontation is received more positively than a more direct and assertive style of confrontation (Czopp et al., 2006; Gervais & Hillard, 2014), others find support for the opposite (Dickter et al., 2012). Previous studies have shown that women who engage in hostile confrontations, such as slapping the perpetrator, are penalized for their behavior in contrast to women who do not (Becker & Barreto, 2014). This penalty for hostility may not necessarily apply to men, because aggression and physical strength are a common way to assert manhood (Bosson, Vandello, Burnaford, Weaver, & Wasti, 2009). Thus, we also explored whether a more aggressive (vs. polite) style of confrontation may have a positive impact on societal perceptions of male confronters.

5.1. Method

5.1.1. Participants

362 students at the University of Osnabrück and residents of the city of Osnabrück participated in the study. We excluded data of 76 participants from the analyses: 49 who failed manipulation checks[4] and 27 who either did not fill out the survey completely or were not serious in their responses (i.e., a research assistant indicated that two participants filled out the survey together), which reduced the final sample to 284 individuals (148 Women and 136 Men; M_{age} = 28.48, SD = 12.29; 94.7% participants were German). Majority of the participants were either studying (46.1%) or working (39.8%). The study was administered in German language, in both paper/pencil form and online via Qualtrics survey platform.

We ran a sensitivity power analysis using G*power program (Faul, Erdfelder, Lang, & Buchner, 2007) with the following parameters: alpha

[4] We also checked whether the findings changed when we included all the participants, however this was not the case (for more details please see Supplementary materials).

level 0.05, power 0.80, sample size 284, three degrees of freedom for the numerator and eight groups in total. The analysis yielded a non-centrality parameter $\lambda = 11.06$, with critical F value of 2.64, and an effect size $f = 0.197$ ($\approx \eta_p^2 = 0.037$). We conclude that the study was powered enough to detect a small to moderate effect.

5.1.2. Manipulation

The participants were first presented with an advertisement for a local paintball club depicting a woman in leather bathing costume holding a paintball gun[5] (see Appendix A). They were asked how much they liked it (1 - *Not at all* to 7 – *Very much*), and whether they were familiar with it (Yes/No). Next, the participants were randomly assigned to one of the four conditions (Polite confrontation vs. Aggressive confrontation vs. No confrontation/Silent ally vs. No confrontation/Silent opponent) and read a short story about Philip M. who visited the paintball club. In two confrontation conditions, Philip thought that the advertisement was offensive to women and decided to confront the manager of the store. We manipulated whether he did so in a polite manner (e.g., he asked politely for the advertisement to be removed), or in an aggressive manner (e.g., he screamed at the manager and threatened to tear down the poster himself). In the other two conditions, Philip did not talk to the manager about the advertisement. However, in one condition the participants read that Philip thought that the advertisement was offensive, but said nothing (i.e., we labeled this condition as *Silent ally*), whereas in the other condition they read that Philip did not see the advertisement as being offensive to women, therefore he did not say anything (i.e., we labeled this condition as *Silent opponent*).

5.1.3. Actor evaluations

We asked the participants to evaluate the extent to which Philip behaved in an appropriate manner or he overreacted (eight items), and whether they perceived him as a likeable person or as too emotional (four items). All items were rated a seven point Likert scale (1 – *Not at all* to 7 – *Very much*). Principal axis factor analysis with Oblimin rotation extracted two factors with eigenvalues larger than one (i.e., 3.02 and 4.98) explaining 60.18% variance. Negative and positive items loaded on the separate factors.[6] The two factors were weakly correlated ($r = -0.16$). Using raw scores, we created two scales that captured whether Philip's behavior was seen as appropriate (items: justified, appropriate, moral, honorable, likeable and friendly; $\alpha = 0.86$) or as an overreaction (items: aggressive, hostile, over the top, exaggerated, emotional and dramatic; $\alpha = 0.89$).

5.1.4. Advertisement evaluation

In order to check participants' views of the advertisement, we asked to what extent they thought the poster was funny (adjectives: funny and witty, $r[282] = 0.88$, $p < .001$) and sexist (adjectives: sexist and chauvinistic, $r[272] = 0.35$, $p < .001$).

5.1.5. Manipulation checks

At the end of the survey, participants responded to two manipulation check questions. The first question asked whether Philip thought that the advertisement was sexist (three answer options: Yes/No/He was unsure). The second question asked about Philip's behavior towards the manager (three answer options: He talked to him politely/He screamed/He said nothing).

[5] The slogan "Bock auf Ballern" written on the advertisement can be interpreted in two ways (and it is expected to contain an element of humor): it refers to a desire to shoot something (with a paintball gun) or to have non-romantic sexual intercourse.

[6] Item loadings for each study can be found in the Supplementary materials.

5.1.6. Procedure

Ten research assistants approached students and residents of the city of Osnabrück and asked them to participate in a ten-minute survey about marketing strategies (we did not reveal the true goal of the survey in order to ensure a broader range of views on gender issues). The participants first filled out the background questions: age, gender, nationality, occupation and political orientation (1 - *Left* to 7 - *Right*, $M = 3.31$, $SD = 1.06$). After the participants provided their initial views of the paintball club advertisement, they were randomly assigned to one of the four conditions and asked to evaluate Philip. We debriefed the participants at the end of the survey, and as a reward for their participation offered them the possibility to participate in a lottery and to win a 20 Euro Amazon voucher.

We report all manipulations, exclusion criteria and all the items used in the analyses for each study. We had additional measures and control variables in the questionnaires that can be found in the Supplementary materials. In Study 1, we aimed to collect as large sample as possible (we aimed for minimum of 30 people per cell), because we recruited a community sample in addition to a student sample. We aimed for similar sample sizes in Study 2 and Study 3. Data collection for Study 1 took place between December 2016 and January 2017, for Study 2 March–June 2018, and for Study 3 July 2018 (pilot for Study 3 was conducted in June 2018). Data were analyzed only after data collection was done. All studies have been approved by the ethical committee of the University of Osnabrück.

5.2. Results and discussion

5.2.1. Advertisement evaluation

Prior to running our main analyses, we examined participants' view of the paintball advertisement. Only three participants were familiar with the advertisement. In general, the participants did not like the advertisement ($M = 2.29$, $SD = 1.50$), they thought it was sexist ($M = 4.98$, $SD = 1.45$), and not particularly funny ($M = 2.39$, $SD = 1.61$). However, we found significant gender differences on all three variables suggesting that men had a more positive view of the advertisement than women (see Table 1). We ran the analyses with and without controlling for the advertisement evaluations in the paper, and the findings remained the same.[7] Below, we report the analyses without the covariates.

5.2.2. Actor evaluation

We ran a two-way Multivariate Analysis of Variance with Manipulation (Polite confrontation vs. Aggressive confrontation vs. No confrontation/Silent ally vs. No confrontation/Silent opponent) and Participant Gender (Women vs. Men) as between subject factors with our two dependent variables. The analysis yielded significant multivariate main effects for Participant Gender, Wilks' Lambda = 0.88, $F(2,275) = 18.80$, $p < .001$, $\eta_p^2 = 0.12$, and Behavior, Wilks' Lambda = 0.27, $F(6,552) = 83.75$, $p < .001$, $\eta_p^2 = 0.48$, which were qualified by a significant Participant Gender x Manipulation interaction, Wilks' Lambda = 0.86, $F(6,552) = 7.02$, $p < .001$, $\eta_p^2 = 0.07$. In the second step, we performed univariate ANOVA's, and followed up on the interaction by conducting simple main effect analyses with Bonferroni correction to adjust for multiple comparisons. Means, standard deviations and univariate analyses can be found in Tables 2a and 2b.

First, women's and men's views of the confronters were largely different. Women evaluated polite confrontation, $F(1,276) = 10.76$, $p = .001$, $\eta_p^2 = 0.04$, and aggressive confrontation, $F(1,276) = 21.91$, $p < .001$, $\eta_p^2 = 0.07$, as significantly more appropriate responses than men did. In contrast, men evaluated confrontation as significantly more

[7] The additional analyses including the covariates can be found in the Supplementary materials.

Table 1
Women's and men's evaluation of paintball advertisement.

	Women		Men		*t*-Test
Evaluations	M	SD	M	SD	
Like the advertisement	1.58	0.90	3.06	1.65	$t(204.64) = -9.28$, $p < .001$, $d = 1.30$
Advertisement seen as sexist	5.30	1.36	4.64	1.47	$t(282) = 3.93$, $p < .001$, $d = 0.49$
Advertisement seen as funny	1.73	1.12	3.10	1.77	$t(224.48) = -7.86$, $p < .001$, $d = 1.05$

Table 2a
Women's and men's evaluations of male confronters and non-confronters in Study 1.

Manipulation	Polite confrontation				Aggressive confrontation				No confrontation/silent ally				No confrontation/silent opponent			
Participant Gender	Women (n = 38)		Men (n = 31)		Women (n = 43)		Men (n = 34)		Women (n = 36)		Men (n = 41)		Women (n = 31)		Men (n = 30)	
Evaluations	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Appropriate	4.98^{a}	1.26	4.06^{b}	1.59	3.95^{b}	0.96	2.71^{c}	0.93	3.88^{b}	1.32	3.89^{b}	1.08	2.25^{d}	0.87	3.13^{c}	1.18
Overreaction	2.55^{a}	0.91	3.52^{b}	1.38	4.29^{c}	1.16	5.77^{d}	0.80	2.06^{a}	0.99	2.27^{a}	0.79	1.87^{a}	0.81	2.14^{a}	1.03

Note. Different superscripts denote significant mean differences.

Table 2b
Study 1: ANOVA summary.

	Appropriate				Overreaction			
	F	*df*	*p*	η_p^2	*F*	*df*	*p*	η_p^2
Participant Gender	5.41	1, 276	.021	0.02	37.69	1, 276	< .001	0.12
Manipulation	29.56	3, 276	< .001	0.24	141.24	3, 276	< .001	0.61
Participant Gender * Manipulation	11.40	3, 276	< .001	0.11	6.76	3, 276	< .001	0.07

of an overreaction than women did, even when it was done politely, $F(1,276) = 16.04$, $p < .001$, $\eta_p^2 = 0.06$, and especially when it was done aggressively, $F(1,276) = 42.07$, $p < .001$, $\eta_p^2 = 0.13$.

Second, women and men evaluations of a non-confronter who thought that the advertisement was sexist (i.e., silent ally) did not differ significantly, $F_{appropriate}(1,276) < 0.001$, $p = .994$, $\eta_p^2 < 0.001$. On the other hand, women judged a non-confronter who did not think that the advertisement was sexist (i.e., silent opponent) as behaving significantly less appropriately than men did, $F(1,276) = 8.67$, $p = .004$, $\eta_p^2 = 0.03$.

Third, we explored whether acting had benefits over endorsing gender equality beliefs by comparing a polite confronter to a non-confronter who perceived the advertisement as sexist, but said nothing (i.e., silent ally). Women perceived a polite confronter as behaving significantly more appropriately than a silent ally, $p < .001$, 95%Cl (0.38, 1.81), and they did not think that he overreacted significantly more than a silent ally, $p = .211$, 95%Cl (−0.13, 1.11). Male audience, on the other hand, did not think that a polite confronter acted significantly more appropriately than a silent ally, $p > .99$, 95% CL (−0.56, 0.90), rather men thought that he overreacted significantly more than a silent ally, $p < .001$, 95%Cl (0.62, 1.88).

Allies' actions and attitudes seem to matter to the general audience, albeit to a different extent. For women, acting upon gender equality beliefs was seen as a more appropriate behavior than staying silent in the face of injustice. In contrast, men approved of fellow group members who hold feminist attitudes, but they did not give them credits for their actions. Quite the opposite, men were more likely to see confronters as overreacting. In addition, in line with previous research (e.g., Becker & Barreto, 2014; Czopp et al., 2006), using a more aggressive style of confrontation was met with discontentment even among women, and especially among men. Altogether, men's responses to allies provide initial support for the hypothesis that allies who confront discrimination may not necessarily be evaluated more positively than those who stay silent by their fellow group members. However, it is not possible to say whether men do not think that allies should be praised for confronting sexism or may not generally approve of anyone who confronts sexism. We address this question in Study 2.

6. Study 2

In Study 2, we manipulated the group membership of the confronter (i.e., male confronter vs. female confronter). We reasoned that if our first hypothesis is correct, than a confrontation by a male ally should be perceived as a more appropriate response and less of an overreaction than a confrontation by a female target by everyone. If our second hypothesis is correct, then ally's confrontation should be evaluated more positively (i.e., as more appropriate and less of an overreaction) than a target's confrontation by women, whereas the opposite should hold for men. Moreover, we also included a no-confrontation condition to be able to examine whether the confronters incur costs or benefits for their actions in contrast to those who perceive the norm violation, but do not act (e.g., silent allies). This enabled us to further test whether the audience is more likely to derogate the confronters with whom they share group membership. If the second hypothesis is true this means that ingroup confronters should not be necessarily rewarded and evaluated more positively when they act, but will be more likely seen as

Table 3a
Women's and men's evaluations of ally and target confronters and non-confronters in Study 2.

Behavior	Silence								Confrontation							
Actor	Target				Ally				Target				Ally			
Participant Gender	Women		Men		Women		Men		Women		Men		Women		Men	
	(n = 37)		(n = 36)		(n = 41)		(n = 33)		(n = 51)		(n = 35)		(n = 46)		(n = 36)	
Variables	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Appropriate	4.39	0.88	4.18	1.12	4.19	1.16	4.03	0.97	4.75	1.25	4.25	1.57	5.38	0.91	4.02	1.32
Overreaction	2.21	0.82	2.99	1.38	2.35	1.08	2.81	0.90	3.46	1.22	3.50	1.26	2.71	0.95	3.99	1.23
Support	4.53	1.08	3.76	1.22	4.28	1.24	3.85	1.22	4.44	1.83	4.08	1.67	5.25	1.32	3.55	1.68

overreacting than ingroup non-confronters. In contrast, outgroup confronters should be seen as reacting more appropriately and overreacting less than outgroup non-confronters. Lastly, we explicitly asked the audience how much they supported the confronter or the non-confronter.[8]

6.1. Method

6.1.1. Participants

331 students and residents of the city of Osnabrück participated in the study. We excluded data of 16 participants from the analyses: 15 who failed manipulation checks and one participant who was younger than 18, and thus not eligible to participate in the study without parental consent. The final sample consisted of 175 women, 140 men (M_{age} = 25.21, SD = 7.59; 95.6% participants were German). The majority of the participants were full time students (76.8%) at the University of Osnabrück. The study was administered in German language using the Unipark survey platform. The sample was on average more left than right wing in their political orientation (1 - *Left* to 7 - *Right*, M = 3.28, SD = 1.07).

Sensitivity power analysis using G*power program with the following parameters (alpha level 0.05, power 0.80, sample size 315, one degree of freedom for the numerator and eight groups in total) yielded the following results: noncentrality parameter λ = 7.90, with critical F value of 3.87, and an effect size f = 0.158 ($\approx \eta_p^2$ = 0.024). Thus, the study had enough power to detect a small effect.

6.1.2. Manipulation

In contrast to Study 1, we now manipulated the bystander's group membership (Actor: Target vs. Ally) and whether she/he acted or not (Behavior: Confrontation vs. Silence). As in Study 1, the participants were first presented with the advertisement for a local paintball club and asked how much they liked it (1 - *Not at all* to 7 – *Very much*) and whether they were familiar with it (Yes/No). Next, the participants read a short story about Anna M. (target) or Philip M. (ally) who thought the advertisement was offensive to women. We only manipulated whether target/ally decided to confront the manager of the store in a polite way or remained silent (see Appendix B for exact wording).

[8] One of the reasons why allies may be liked more than targets according to some literature (e.g., Dickter et al., 2012; Drury & Kaiser, 2014), is because the general audience may believe that allies' confrontations will indeed have a more positive effect on the perpetrators than targets' confrontations. Therefore, in Study 2 and Study 3 we also asked the participants whether they believed that confrontation is an effective tool in reducing discrimination. The participants responded to a set of six items asking how likely it is that target/ally will change the manager's opinion or will make him angry. Overall, the participants did not believe that a confrontation would be very successful. Moreover, contrary to Drury and Kaiser's (2014) expectations, allies were not deemed as more successful than targets in changing perpetrator's behavior. This suggests that societal approval for those who confront discrimination is less dependent on instrumental concerns. For more details, please see Supplementary materials.

6.1.3. Actor evaluation

We extended the scales used in Study 1 by incorporating four additional items (i.e., respectable, admirable, complainer and troublemaker; we changed the item dramatic to sensitive) used in previous studies (e.g., Kaiser & Miller, 2001). Principal factor analysis with Oblimin rotation extracted three factors with eigenvalues larger than one (1.10, 2.38 and 6.89) explaining 56.55% variance. The loadings on the third factor were all lower than < 0.46, hence we decided to go with the two factor solution that explained 52.37% variance. As in Study 1, negative and positive items loaded on the separate factors. Using raw scores, we created the scale that measured whether the behavior was seen as appropriate (eight items, α = 0.89) or as an overreaction (eight items; α = 0.87). All items were asked on a seven point Likert scale (1 – *Strongly disagree* to 7 – *Strongly agree*). The two variables were negatively correlated, $r(313) = -0.52$, $p < .001$.

6.1.4. Support

The participants were asked to what extent they agreed (1 – *Strongly disagree* to 7 – *Strongly agree*) with the target's or ally's opinion, her/his behavior and perceived her/him as a role model. We combined these three items as an indication of participants' support (α = 0.81). Perceived support was strongly positively correlated with perceptions of appropriateness, $r(313) = 0.82$, $p < .001$, and negatively correlated with perceptions of overreaction, $r(313) = -0.60$, $p < .001$.

6.1.5. Manipulation checks

At the end of the survey, the participants responded to two manipulation check questions whether they remembered who they were asked to evaluate (two answer options: Anna vs. Philip) and what the person did (two answer options: She/He talked to the manager politely vs. She/He said nothing). The procedure was the same as in the previous study.

6.2. Results and discussion

6.2.1. Advertisement evaluation

Only five participants were familiar with the advertisement. Similar to Study 1, the participants did not like the advertisement (M = 2.36, SD = 1.52), though men liked it significantly more than women did (M = 3.19, SD = 1.68 vs. M = 1.69, SD = 0.96), $t(209.12) = 9.37$, $p < .001$, d = 1.30. We ran the analyses with and without controlling for the advertisement evaluation and the findings remained the same. Below, we report the analyses without the covariate (see Supplementary materials for analysis with the covariate).

6.2.2. Actor evaluation

A three-way MANOVA with Participant Gender (Women vs. Men), Actor (Ally vs. Target) and Behavior (Confrontation vs. Silence) with two dependent variables yielded the following: a significant main effect of Participant Gender, Wilks' Lambda = 0.92, $F(2,306) = 13.82$, $p < .001$, $\eta_p^2 = 0.08$, and Behavior, Wilks' Lambda = 0.73, F

Table 3b
Study 2: ANOVA summary.

	Appropriate				Overreacting				Support			
	F	*df*	*p*	η_p^2	*F*	*df*	*p*	η_p^2	*F*	*df*	*p*	η_p^2
Participant Gender	17.58	1, 307	< .001	0.05	25.43	1, 307	< .001	0.08	24.67	1, 307	< .001	0.07
Actor	0.01	1, 307	.926	0.00	0.33	1, 307	.565	0.00	0.03	1, 307	.855	0.00
Behavior	9.00	1, 307	.003	0.03	41.59	1, 307	< .001	0.12	1.80	1, 307	.181	0.01
Participant Gender ∗ Actor	2.38	1, 307	.124	0.01	3.32	1, 307	.069	0.01	2.32	1, 307	.129	0.01
Participant Gender ∗ Behavior	7.91	1, 307	.005	0.03	0.02	1, 307	.881	0.00	1.68	1, 307	.196	0.01
Actor ∗ Behavior	1.99	1, 307	.159	0.01	0.19	1, 307	.668	0.00	0.44	1, 307	.508	0.00
Participant Gender ∗ Actor ∗ Behavior	2.97	1, 307	.086	0.01	9.32	1, 307	.002	0.03	6.47	1, 307	.011	0.02

(2,306) = 55.85, $p < .001$, $\eta_p^2 = 0.27$, which were qualified by a significant two-way Participant Gender x Behavior interaction, Wilks' Lambda = 0.96, $F(2,306) = 5.66$, $p = .004$, $\eta_p^2 = 0.04$, and a significant three-way interaction, Wilks' Lambda = 0.97, $F(2,306) = 4.65$, $p = .01$, $\eta_p^2 = 0.03$. Complete MANOVA output can be found in the Supplementary materials. Means, standard deviations and univariate analyses are reported in Tables 3a and 3b.

Our analytic strategy was twofold: First, we performed a series of simple main effect analyses to test whether an ally confronter was evaluated differently than a target confronter separately for men and women. These were our primary tests examining the evidence for the two hypotheses. Second, we performed an additional test of *Hypothesis 2* and compared the participants' perceptions of confronters and non-confronters. The goal of this additional test was to examine whether ingroup members were evaluated less positively when they acted as opposed when they stayed silent, whereas outgroup members were not (or to a less extent).

6.2.2.1. Primary test: ally vs. target confronter. First, women evaluated target's confrontation as significantly less appropriate behavior than ally's confrontation, $F(1,307) = 7.12$, $p = .008$, $\eta_p^2 = 0.02$. At the same time, women perceived a target confronter as overreacting to a significantly higher degree than an ally confronter, $F(1,307) = 10.83$, $p = .001$, $\eta_p^2 = 0.03$. In contrast, men's views of the two confronters did not significantly differ in terms of appropriateness, $F(1,307) = 0.71$, $p = .399$, $\eta_p^2 < 0.001$. However, men thought that the ally was overreacting more than the target, although the difference was not significant, $F(1,307) = 3.44$, $p = .065$, $\eta_p^2 = 0.01$. The findings are displayed in Fig. 1.

Additionally, although we did not a priori hypothesize about this, we note that women's and men's evaluations of the ally confronter were significantly different: women perceived ally's confrontation as more appropriate behavior than men, $F(1,307) = 27.59$, $p < .001$, $\eta_p^2 = 0.08$, and less of an overreaction than men, $F(1,307) = 26.51$, $p < .001$, $\eta_p^2 = 0.08$. In contrast, women's and men's views of the target confronter were not significantly different: $F_{appropriate}(1,307) = 3.76$, $p = .053$, $\eta_p^2 = 0.01$, $F_{overreaction}(1,307) = 0.03$, $p = .869$, $\eta_p^2 < 0.001$.

6.2.2.2. Additional test: confronters vs. non-confronters. In line with *Hypothesis 2*, women did not think the target acted more appropriately when she confronted than when she remained silent, $F(1,307) = 1.99$, $p = .159$, $\eta_p^2 = 0.01$, rather she was seen as a complainer to a significantly higher degree when she confronted than when she remained silent, $F(1,307) = 26.56$, $p < .001$, $\eta_p^2 = 0.08$. In contrast, women thought the ally acted more appropriately when he confronted than when he remained silent, $F(1,307) = 22.47$, $p < .001$, $\eta_p^2 = 0.07$, and they did not think he was overreacting more when he confronted as opposed to when he did not, $F(1,307) = 2.23$, $p = .137$, $\eta_p^2 = 0.01$.

On the other hand, men did not think that the ally acted more appropriately when he confronted than when he remained silent, $F(1,307) = 0.004$, $p = .953$, $\eta_p^2 < 0.001$. Rather, men perceived the ally as overreacting more when he acted than when he remained silent, $F(1,307) = 18.96$, $p < .001$, $\eta_p^2 = 0.06$. Moreover, men did not think that the target was acting more appropriately when she confronted as opposed to when she did not, $F(1,307) = 0.06$, $p = .810$, $\eta_p^2 < 0.001$, but she was also not perceived as overreacting more when she confronted as opposed to when she did not, $F(1,307) = 3.64$, $p = .058$, $\eta_p^2 = 0.01$.

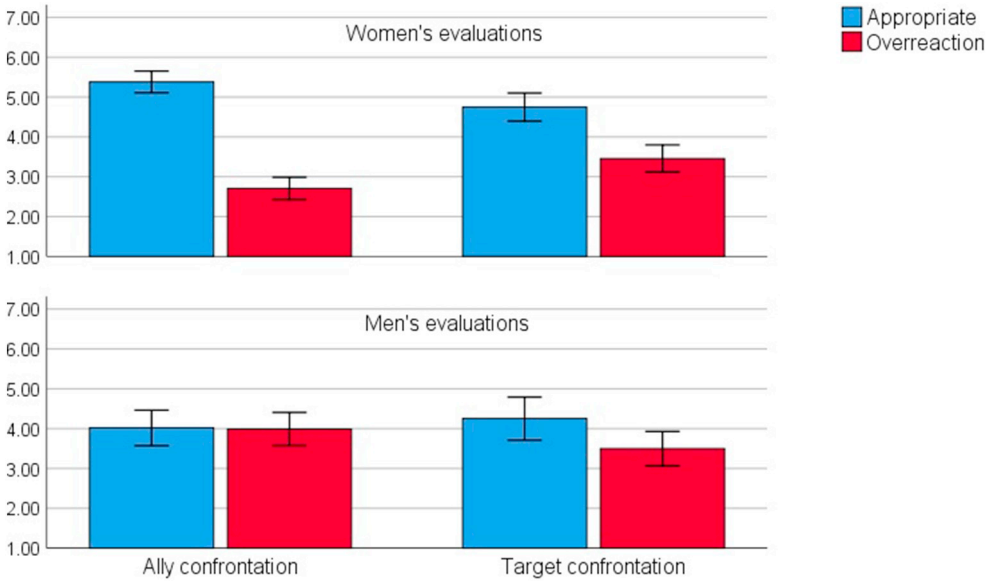

Fig. 1. Women's and men's evaluations of ally and target confrontation. Error bars represent standard errors.

6.2.3. Support

A univariate analysis of variance on support yielded a significant main effect of Participant Gender and a significant three-way interaction (for more details please see Tables 3a and 3b). Importantly, women supported the target confronter significantly less than the ally confronter, $F(1,307) = 7.56$, $p = .006$, $\eta_p^2 = 0.02$. In contrast, the means point in the direction that men supported the ally confronter, less than the target confronter, albeit this difference was not significant, $F(1,307) = 2.38$, $p = .124$, $\eta_p^2 = 0.01$.

Additionally, men supported the ally confronter significantly less than women did, $F(1,307) = 27.90$, $p < .001$, $\eta_p^2 = 0.08$, whereas women's and men's support for the target confronter did not differ significantly, $F(1,307) = 1.30$, $p = .255$, $\eta_p^2 < 0.001$.

In sum and in contrast to *Hypothesis 1*, allies were not unequivocally liked more than targets. Instead, an ally's confrontation was seen as a more appropriate response and supported more than a target's confrontation by members of disadvantaged group. At the same time, members of the advantaged group were somewhat less supportive of confrontation when it was done by an ally as opposed to when it was done by a target, rather they perceived it as somewhat more of an overreaction. Moreover, the comparison between confronters and non-confronters suggested that a confrontation by an outgroup member was less likely to be dismissed as an overreaction than a confrontation by an ingroup member, supporting *Hypothesis 2*.

One can speculate that the participants may have distanced from their fellow group members who confronted, because they violated prescriptive norms and stereotypes associated with being a "good" woman (e.g., warm and caring) vs. being a "good" man (e.g., macho; Fiske & Stevens, 1993). Previous research found that those who violate gender stereotypes face backlash (Phelan, Moss-Racusin, & Rudman, 2008; Rudman et al., 2013). If the stereotype violation had occurred, one would expect confronters to be perceived as less prototypical of their gender group than non-confronters. We explored this explanation by examining the extent to which participants perceived target's or ally's confrontation as prototypical for men and women (please see Supplementary materials for more details). Overall, neither men nor women perceived their fellow group members who confronted as less prototypical than those who stayed silent, which suggests that stereotype violation could not be the only process driving the effects.

7. Study 3

Study 3 aimed to replicate the findings of Study 2 in the context of racism and discrimination against Black Americans. First, we ran a short pilot study ($N = 50$) to find an advertisement that was perceived as offensive due to its derogatory portrayal of Black Americans (see Supplementary materials for more details). We adjusted the manipulation by keeping the gender of the actor constant (i.e., man) and only changing the racial background of the confronter (Black American or White American). Moreover, target confronters are often dismissed because they are seen as impolite (Dodd et al., 2001; Swim & Hyers, 1999), and we explored whether the same applies to allies.

7.1. Method

7.1.1. Participants

535 Amazon Mechanical Turk workers participated in the study for a small monetary reward 0.50$. Due to recent concerns over the quality of the data obtained through Amazon Mechanical Turk[9] (Hauser, Paolacci, & Chandler, 2018), we included both manipulation and attention checks and screened the data for unusual responses. This resulted in data of 154 participants being excluded from the analyses: 37 participants who did not finish the survey, 103 who failed manipulation checks,[10] 14 participants who did not fill out the survey seriously (they finished the survey in less than 3 min, their replies on the open questions at the end of the survey were not interpretable, and one person responded only using the midpoint of the scale). The final sample was thus reduced to 381 individuals (188 Black Americans, 193 White Americans; $M_{age} = 36.44$, $SD = 12.03$). The study was administered in English language using the Qualtrics survey platform. The sample was on average leaning towards more liberal political orientation (1 - *Liberal* to 7 - *Conservative*, $M = 3.35$, $SD = 1.75$); participants on average self-identified more strongly as a Democrat ($M = 4.11$, $SD = 2.13$) or Independent ($M = 4.05$, $SD = 2.11$), and less strongly as a Republican ($M = 2.56$, $SD = 1.99$).

Sensitivity power analysis using G*power program with the following parameters (alpha level 0.05, power 0.80, sample size 381, one degree of freedom for the numerator and eight groups in total) yielded the following results: noncentrality parameter $\lambda = 7.89$, with critical F value of 3.87, and an effect size $f = 0.14$ ($\approx \eta_p^2 = 0.02$). Thus, the study had enough power to detect a small to moderate effect.

7.1.2. Manipulation

We used the same design as in Study 2. The participants were first presented with an advertisement for a cosmetic product (the advertisement depicted a Black man throwing away his African-American mask with the slogan "Re-civilize yourself") and asked how much they like it (1 - *Not at all* to 7 – *Very much*), and whether they were familiar with it (Yes/No). Next, the participants read a short story about Michael C., a Black American vs. White American student who saw the advertisement in his local drugstore. In all conditions, the actor thought that the advertisement was racist and either politely asked the manager to remove it or not (please see Appendix C for exact wording).

7.1.3. Dependent variables

We used the items from Study 2. Principle axis factors with Oblimin rotation on evaluation items extracted two factors explaining 65.15% variance with eigenvalues larger than one (3.78 and 7.32). Positive and negative items loaded on separate factors. We calculated the scales based on the original scores for appropriate reaction (eight items, $\alpha = 0.94$), and overreaction (eight items, $\alpha = 0.94$). We also measured participants' support for confronters and non-confronters as in Study 2 (three items, $\alpha = 0.77$). Lastly, we added a measure of perceived impoliteness and asked the participants to what extent they thought that Michael: is a polite person/does not care if he offends someone else/does not mind violating politeness norms (we recoded the first item so the higher scores indicate more impoliteness, $\alpha = 0.73$). The correlations between the dependent variables can be found in Table 4.

7.1.4. Manipulation and attention checks

At the end of the survey, the participants responded to two manipulation check questions whether they remembered Michael's racial background (two answer options: White American/Black American) and his behavior (two answer options: He talked to the manager politely/He said nothing). Moreover, we included an attention check. The participants read the following: "*What is the topic of the study? Sometimes participants do not carefully read the instructions. In order to correctly*

[9] We also encountered 69 entries (and deleted them) with duplicate IP addresses.

[10] The largest number of mistakes on manipulation check questions was made in the condition where a White American character did not confront (55.3%), and we had to oversample participants for this condition (we screened the data to see whether we had enough people who passed manipulation checks, we did not test our hypotheses). Those who made a mistake, thought that the character was a Black American (and the means in this conditions were somewhat higher in the full sample). Because we excluded quite a substantive number of people, we also ran the analyses on the full dataset. The key differences in the evaluations of ally and target confronters remained the same. We report the descriptive statistics and analyses in the Supplementary materials.

Table 4
Correlations between variables in Study 3.

	Appropriate	Overreaction	Support
Overreaction	−0.32**		
Support	0.80**	−0.50**	
Impoliteness	−0.21**	0.61**	−0.27**

** $p < .001$.

answer this question and earn your credits, please select the option Other and write down the name of your favorite movie". We included five answer options: (1) advertisement vs. (2) racial issues vs. (3) police violence vs. (4) I cannot remember vs. (5) other. Because the participants had to ignore the obvious answers, this question enabled us to discern between the participants who thoroughly read our questions and those who were just skipping to the answer choices (Hauser, Ellsworth, & Gonzalez, 2018). We included the participants who either answered all questions correctly or made only one mistake (i.e., failed the attention check, but correctly answered both manipulation checks).

7.2. Results and discussion

7.2.1. Advertisement evaluations

Twenty-one participants said they knew the advertisement. Similar to Studies 1 and 2, participants did not like the advertisement ($M = 2.12$, $SD = 1.61$), but their views were not significantly different, $t(379) = -1.29$, $p = .198$, $d = 0.13$.

7.2.2. Evaluations

A three-way MANOVA with Participant Race (Black Americans vs. White Americans), Actor (Ally vs. Target), and Behavior (Confrontation vs. Silence) yielded significant main effects of Participant Race, Wilks' Lambda = 0.97, $F(2,372) = 5.35$, $p = .005$, $\eta_p^2 = 0.03$, Actor Race, Wilks' Lambda = 0.98, $F(2,372) = 4.51$, $p = .012$, $\eta_p^2 = 0.02$, and Behavior, Wilks' Lambda = 0.74, $F(2,372) = 66.14$, $p < .001$, $\eta_p^2 = 0.26$. The interactions were not significant. Complete MANOVA can be found in the Supplementary materials. Means, standard deviations and univariate analyses are reported in Tables 5a and 5b.

This study was designed as a replication of Study 2, and we used the same analytic strategy and performed the planned contrasts. More specifically, first we compared Black Americans and White Americans evaluations of the two confronters (i.e., primary tests for the two hypotheses). Second, we compared whether ingroup members were evaluated as overreacting more when they acted as opposed when they stayed silent, whereas outgroup confronters were not (or to a less extent), as our additional test in line with Hypothesis 2.

7.2.2.1. Primary test: ally vs. target confronters. In contrast to Study 2, Black Americans' views of the two confronter did not differ significantly: they did not think that the ally acted more appropriately than the target confronter, $F(1,373) = 1.14$, $p = .286$, $\eta_p^2 = 0.003$, nor they perceived the target as overreacting more than the ally, $F(1,373) = 0.25$, $p = .615$, $\eta_p^2 = 0.001$. On the other hand, White Americans perceived both target's and ally's confrontation as equally appropriate, $F(1,373) = 3.50$, $p = .062$, $\eta_p^2 = 0.01$, but they evaluated a confrontation by an ally as significantly more of an overreaction than a confrontation by a target, $F(1,373) = 7.44$, $p = .007$, $\eta_p^2 = 0.02$. The findings are displayed in Fig. 2.

Additionally, like in Study 2, Black Americans' views and White Americans' views of the target confronter did not differ significantly: $F_{appropriate}(1,373) = 1.16$, $p = .281$, $\eta_p^2 = 0.003$, $F_{overreaction}(1,373) = 0.07$, $p = .797$, $\eta_p^2 < 0.001$. In contrast, participants' views of the ally confronter diverged significantly: compared to White Americans, Black Americans perceived the ally as acting more appropriately, $F(1,373) = 16.63$, $p < .001$, $\eta_p^2 = 0.04$, and as overreacting

Table 5a
Black Americans' and White Americans' evaluations of confronters and non-confronters in Study 3.

Behavior	Silence						Confrontation					
Actor	Target			Ally			Target			Ally		
Participant race	Black Americans		White Americans	Black Americans		White Americans	Black Americans		White Americans	Black Americans		White Americans
	(*n* = 48)		(*n* = 49)	(*n* = 44)		(*n* = 45)	(*n* = 48)		(*n* = 47)	(*n* = 48)		(*n* = 52)
Variables	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Appropriate	4.85	1.08	4.72	1.59	4.26	1.53	3.99	1.24	5.48	1.45	5.18	1.10	5.77	1.10	4.68	1.48
Overreaction	1.87	0.95	2.00	1.14	2.17	1.46	2.30	1.15	2.86	1.73	2.79	1.28	2.72	1.52	3.54	1.52
Support	5.19	0.98	4.90	1.39	4.39	1.22	4.12	1.23	5.40	1.77	5.18	1.37	5.73	1.38	4.27	1.77
Impoliteness	2.48	1.25	2.48	1.17	2.53	1.33	2.14	0.98	3.51	1.53	3.10	1.38	3.27	1.49	3.82	1.48

Table 5b
Study 3: ANOVA summary.

	Appropriate				Overreaction			
Factors	*F*	*df*	*p*	η_p^2	*F*	*df*	*p*	η_p^2
Participant Race	10.54	1,373	.001	0.03	3.19	1,373	.075	0.01
Actor	7.92	1,373	.005	0.02	4.65	1,373	.032	0.01
Behavior	35.75	1,373	< .001	0.09	40.42	1,373	< .001	0.10
Participant Race * Actor	2.87	1,373	.091	0.01	2.52	1,373	.113	0.01
Participant Race * Behavior	3.28	1,373	.071	0.01	0.77	1,373	.380	0.00
Actor * Behavior	4.17	1,373	.042	0.01	0.00	1,373	.986	0.00
Participant Race * Actor * Behavior	1.45	1,373	.229	0.00	2.53	1,373	.113	0.01
	Support				**Impoliteness**			
Factors	*F*	*df*	*p*	η_p^2	*F*	*df*	*p*	η_p^2
Participant Race	14.81	1,372	< .001	0.04	0.21	1,373	.649	0.00
Actor	13.90	1,372	< .001	0.04	0.12	1,373	.734	0.00
Behavior	11.48	1,372	.001	0.03	54.76	1,373	< .001	0.13
Participant Race * Actor	4.33	1,372	.038	0.01	1.08	1,373	.300	0.00
Participant Race * Behavior	3.75	1,372	.054	0.01	0.90	1,373	.344	0.00
Actor * Behavior	3.02	1,372	.083	0.01	1.96	1,373	.163	0.01
Participant Race * Actor * Behavior	4.65	1,372	.032	0.01	6.09	1,373	.014	0.02

less, $F(1,373) = 8.96$, $p = .003$, $\eta_p^2 = 0.02$.

7.2.2.2. Additional test: confronters vs. non-confronters. Specific contrasts indicated that Black Americans thought that the target acted more appropriately when he confronted the advertisement than when he remained silent, $F(1,373) = 5.21$, $p = .023$, $\eta_p^2 = 0.01$. However, Black Americans perceived the target as overreacting more when he confronted than when he remained silent, $F(1,373) = 12.56$, $p < .001$, $\eta_p^2 = 0.03$. Replicating Study 2, Black Americans evaluated the ally as acting more appropriately when he confronted as opposed to when he remained silent, $F(1,373) = 29.43$, $p < .001$, $\eta_p^2 = 0.07$, without judging the confrontation as an overreaction in comparison to staying silent, $F(1,373) = 3.69$, $p = .056$, $\eta_p^2 = 0.01$.

White Americans thought that the ally acted more appropriately when he confronted than when he remained silent, $F(1,373) = 6.36$, $p = .012$, $\eta_p^2 = 0.02$, but as in Study 2 they also thought that he overreacted to a significantly higher degree when he confronted as opposed to when he remained silent, $F(1,373) = 19.86$, $p < .001$, $\eta_p^2 = 0.05$. The target confronter was not seen as acting more appropriately than a target non-confronter, $F(1,373) = 2.80$, $p = .095$, $\eta_p^2 = 0.01$, and even though the target confronter was also perceived as overreacting more than the non-confronter, $F(1,373) = 7.99$, $p = .005$, $\eta_p^2 = 0.02$, this effect was smaller in comparison to the ally evaluation.

7.2.3. Support

A univariate analysis of variance yielded all three significant main effects, a significant two-way Participant Race × Actor interaction and a significant three–way interaction (please see Tables 5a and 5b). First, in contrast to Study 2, Black Americans supported the ally and the target confronter to the same extent, $F(1,372) = 1.27$, $p = .261$, $\eta_p^2 < 0.001$. On the other hand and in line with Hypothesis 2, White Americans supported the target confronter significantly more than the ally confronter, $F(1,372) = 10.10$, $p = .002$, $\eta_p^2 = 0.03$. In addition, Black Americans' and White Americans' support for the target confronter did not differ significantly, $F(1,372) = 0.60$, $p = .439$, $\eta_p^2 < 0.001$, whereas White Americans were significantly less supportive of the ally than Black Americans, $F(1,372) = 26.16$, $p < .001$, $\eta_p^2 = 0.07$.

7.2.4. Impoliteness

A univariate analysis of variance yielded a significant main effect of Behavior and a significant three–way interaction. Importantly, Black Americans did not perceive the target confronter as more impolite than the ally confronter, $F(1,373) = 0.79$, $p = .375$, $\eta_p^2 < 0.001$. White Americans, on the other hand, thought that the ally confronter was more impolite than the target confronter, $F(1,373) = 7.14$, $p = .008$, $\eta_p^2 = 0.02$. Again, White Americans perceived the ally as more impolite than Black Americans did, $F(1,373) = 4.20$, $p = .04$, $\eta_p^2 = 0.01$, whereas their views of the target were not significantly different, $F(1,373) = 2.27$, $p = .133$, $\eta_p^2 = 0.01$.

In contrast to Study 2, members of the disadvantaged group had similarly positive views about ally and target confronters. Replicating Study 2 findings, members of advantaged group had a less positive view of the ally compared to the target confronter, and were less likely to support the ally's action. One possible explanation of this effect is that White Americans responded in a socially desirable way when they were asked to evaluate a target confronter. Research on aversive racism (Dovidio & Gaertner, 2004) found that liberal audiences' preferential treatment of Black Americans over White Americans can also be driven by prejudicial attitudes (Nail, Harton, & Decker, 2003). To address this possible explanation, we explored whether political orientation moderated White Americans' perceptions of ally and target confronters. However, we did not find a significant interaction on any of our dependent variables (for more details see Supplementary materials). Thus, although aversive racism may have played some role, this alone cannot explain the observed effects.

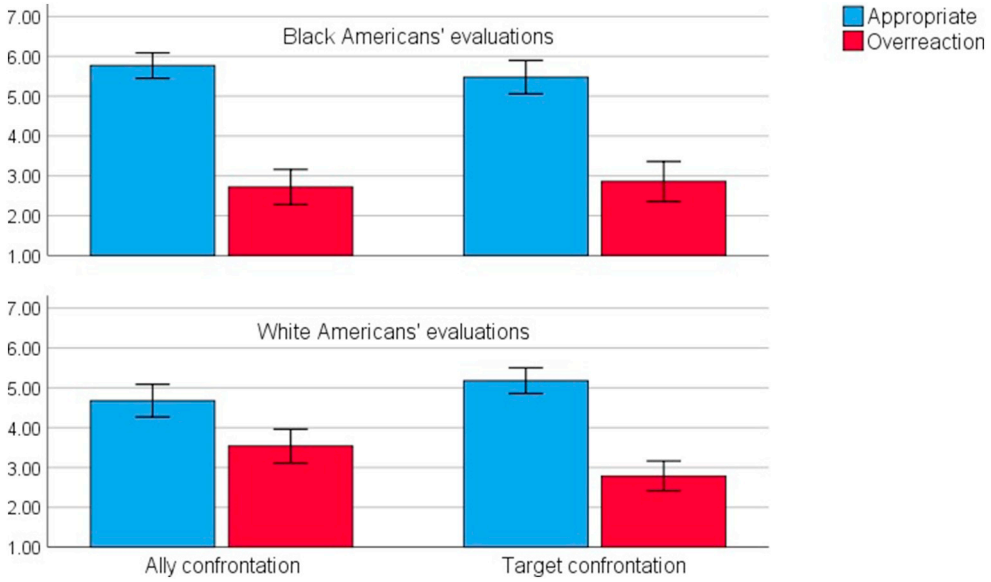

Fig. 2. Black Americans' and White Americans' evaluations of confrontation. Error bars represent standard errors.

8. General discussion

Are allies like Ronan Farrow who engage in potentially morally courageous behavior supported more than Jodi Kantor and Megan Twohey? Our answer to this question is that it depends on the audience. Prior work on confrontation suggested that allies in contrast to targets receive more societal support when they confront discrimination (*Hypothesis 1*). Using the insights from the literature on do-gooder derogation (Monin, 2007; Monin et al., 2008), we proposed that allies and targets risk being evaluated negatively primarily by their fellow group members.

In Study 1 we explored and found that members of advantaged group, in contrast to members of disadvantaged group, were more likely to perceive confrontation by an ally as an overreaction and especially if it was done aggressively. In Studies 2 and 3, we tested our two competing hypotheses by looking at the societal support for allies and targets who confronted sexist or racist behavior. Overall, we did not find much evidence to support *Hypothesis 1* for the reason that allies were not unequivocally perceived more positively than targets. The results were more consistent with *Hypothesis 2*: allies and targets received a lot of support, but also risked more backlash (i.e., in terms of being perceived as overreacting) from ingroup members than outgroup members. However, we also identified important contextual differences, which to some extent limit the support for *Hypothesis 2*. In the context of sexism, members of the disadvantaged group preferred the ally confronter more strongly than the target confronter (Study 2), but this effect did not replicate in the context of racism (Study 3). Members of the advantaged group thought the ally confronter was more overreacting than the target confronter. But, we note that there was only a tendency for this effect in the context of sexism (Study 2), whereas we found clearer support for this effect in the context of racism (Study 3).

Additionally, Study 2 and Study 3 revealed that members of advantaged and disadvantaged groups had similar perceptions of target confronters. In contrast, allies were supported more by members of disadvantaged groups than advantaged groups. These findings suggest that societal perceptions of allies who engage in confrontation of discrimination may be more divided than societal perceptions of targets. We discuss the implications of our findings below.

8.1. Theoretical implications

This paper contributes to the literature on moral courage by showing how interventions against norm violations are perceived by the society at large. Prior research suggested that bystanders who intervene fear and expect societal reprisal (Greitemeyer et al., 2007). On a positive side, our studies seem to suggest that the general audience is rather supportive of those who act, and perhaps this knowledge can be used to motivate people to confront injustice. Nevertheless, this may be due to stronger societal norms against sexism or racism in the populations we examined, and may not necessarily apply to all bystander interventions.

At the same time, fear of societal reprisal is not without its grounds. Namely, those who acted to protect societal norms against discrimination were also perceived as overreacting to an extent. Our findings however contradict previous assumptions that targets are more prone to societal backlash (Dickter et al., 2012; Drury & Kaiser, 2014), and showcase that allies are not immune to it either. Even though the absolute ratings on perceptions of overreaction and impoliteness were not high in our studies (i.e., below scale midpoint), reputational costs may still demotivate people from acting. We suspect that this may be particularly the case for people high in trait social anxiety who report lower intentions to engage in morally courageous behavior (Baumert et al., 2013), and who worry more about losing public support.

Our work poses new and intriguing questions about the processes that may lead to societal (dis)approval of morally courageous individuals. According to Monin (2007), the reason why the audience may distance from moral exemplars is due to perceived threats to individuals' moral identities. However, our findings suggest that there may be several different processes at play. For instance, women's more enthusiastic support for the ally confronter in contrast to the target

confronter may be driven by women's stronger internalization of negative attitudes towards feminists (Anastosopoulos & Desmarais, 2015; Cottrell & Neuberg, 2005), and/or endorsement of benevolent sexism and attraction to men who exhibit characteristics of a "high status protector" (Bohner, Ahlborn, & Steiner, 2010; Glick & Fiske, 2001). Similar attitudes do not apply to the context of racism, which may explain why we did not find the same pattern among Black Americans. Likewise, somewhat weaker pattern among men in Study 2 in contrast to White Americans in Study 3 may be due to stronger norms against racist as opposed to sexist behavior (Cowan & Hodge, 1996; Czopp & Monteith, 2003; Rodin, Price, Bryson, & Sanchez, 1990). Consequently, men might perceive less societal pressure not to appear as sexist, and therefore they may be less likely to see confrontation of sexism as a 'threatening' situation (and by extension the ally as a threatening figure). Thus, future research on do-gooder derogation should take into account societal norms and attitudes that exist in different contexts that may affect the extent to which individuals tend distance from moral exemplars.

Furthermore, the responses to confronters in our studies may have been driven by perceived threats to group status and/or group image as well as to the individual moral identity. Self- categorization theory suggests that threats can be experienced both on the individual as well as on the group level, and individuals can act defensively when their group's status is threatened (Branscombe, Ellemers, Spears, & Doosje, 1999). In Study 2 and Study 3, we included for exploratory reasons the questions about whether the confronters threatened the public image of their group. Although we did not find much evidence that ingroup confronters were seen as more damaging to group's image than outgroup confronters (for more details please see Supplementary materials), a combination of perceived threat to group status as well as to personal image may together result in less positive evaluations of ingroup confronters. One way to examine whether threats play an important role in this process is to use an intervention that should eliminate their effects. For example, Monin et al. (2008) found no evidence of do-gooder derogation when the audience was given the opportunity to self-affirm (for more details please see Study 4, Monin et al., 2008). Future research could investigate whether providing the participants with the opportunity to affirm their self-worth may reduce or eliminate the tendency to perceive the confronters as complainers.

8.2. Limitations and directions for future research

The studies have several limitations. First, the choice of the stimuli limits the external validity of our findings. In all three experiments we only looked at the confrontation of one specific type of discrimination, that is, offensive portrayals of disadvantaged groups in advertising. We chose this context because it is an example of blatant discrimination and therefore eliminates (or at least decreases significantly) the possibility to deny the injustice. Nevertheless, we found that acting against a clear norm violation brings about some reputational costs. In a different context, such as the workplace, discrimination is more likely to be expressed in a subtle manner, which makes it even harder to argue that norms have been violated. In this situation, we would expect that those who dare to speak out may face even more backlash.

Second, our manipulations may be seen as less ecologically valid, as we relied on hypothetical scenarios. A more powerful test of the hypotheses would be to have participants witness someone confronting in real life. However, people rarely engage in confrontation of discrimination (Hyers, 2007; Mallett & Melchiori, 2014). Thus, the general audience is likely to hear through the grapevine or media about such an event (and form an opinion about it), than they are to witness it. Importantly, our findings correspond with the study by Czopp et al. (2006) that used real confrontation in the lab and found Black confronters to be more effective than White confronters in eliciting self-directed negative affect among the perpetrators. Thus, we expect that witnessing confrontation may not cancel, but rather exacerbate the differences in the evaluations of allies and targets.

9. Conclusion

Confrontation is an important way to reduce discrimination, however those who engage in it may not necessarily be seen as heroes. Societal approval for those who confront discrimination depends on the relationship between the confronters and their audience, and sometimes the audience may fail to show solidarity with those who act. We hope that our research sheds new light on these issues and may be used to provide a more supportive community to all who dare to fight injustice.

Appendix A. Materials used in Study 1

Fig. 3. The advertisement used in Studies 1 and Study 2.

Table 6
Manipulations used in Study 1.

Behavior	Philip M. visited the local paintball field and saw the poster shown above….
Polite confrontation	Philip M. was upset about the poster and felt it was disrespectful towards women. He talked to the manager of the place and kindly asked for the poster to be removed.
Aggressive confrontation	Philip M. was angry about the poster and felt it was extremely disrespectful towards women. He yelled at the manager to remove the poster, otherwise he would break it himself.
No confrontation/silent ally	Philip M. was uncertain about the poster and felt it was disrespectful towards women. However, he decided to keep his opinion to himself and said nothing to the manager.
No confrontation/silent opponent	Philip M. had no doubts about the poster and felt it was not disrespectful towards women. He kept his opinion to himself and said nothing to the manager.

Appendix B. Manipulation in Study 2

Table 7
Manipulations used in Study 2.

Behavior	Anna M./Philip M. visited the local paintball club and saw the poster ….
Polite confrontation	**Anna M./Philip M.** was upset about the poster and felt it was disrespectful towards women. **She/He** talked to the manager of the place and kindly asked for the poster to be removed.
Silence	**Anna M./Philip M.** was upset[a] about the poster and felt it was disrespectful towards women. However, **she/he** decided to keep **hers/his** opinion to **her/himself** and said nothing to the manager.

Note.
[a]We replaced uncertain with upset in Study 2, so it matches the confrontation condition.

Appendix C

Fig. 4. The advertisement used in Study 3.

Table 8
Manipulations used in Study 3.

Behavior	
Polite confrontation	Michael C., **a Black American/White American** student, visited a local drugstore in his hometown and saw the advertisement shown before. Michael was upset about the advertisement and felt it was inappropriate and discriminating against Black Americans. He talked to the manager of the store and kindly asked for the advertisement to be removed.
Silence	Michael C., **a Black American/White American** student, visited a local drugstore in his hometown and saw the advertisement shown before. Michael was upset about the advertisement and felt it was inappropriate and discriminating against Black Americans. However, he decided to keep his opinion to himself and said nothing to the manager.

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2019.103832.

References

Abrams, D., Marques, J. M., Bown, N., & Henson, M. (2000). Pro-norm and anti-norm deviance within and between groups. *Journal of Personality and Social Psychology, 78*, 906–912. https://doi.org/10.1037/0022-3514.78.5.906.

Algoe, S. B., & Haidt, J. (2009). Witnessing excellence in action: The 'other-praising' emotions of elevation, gratitude, and admiration. *Journal of Positive Psychology, 4*, 105–127. https://doi.org/10.1080/17439760802650519.

Anastosopoulos, V., & Desmarais, S. (2015). By name or by deed? Identifying the source of the feminist stigma. *Journal of Applied Social Psychology, 45*, 226–242. https://doi.org/10.1111/jasp.12290.

Ariyanto, A., Hornsey, M. J., & Gallois, C. (2010). United we stand: Intergroup conflict moderates the intergroup sensitivity effect. *European Journal of Social Psychology, 40*, 169–177. https://doi.org/10.1002/ejsp.628.

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

Baumert, A., Halmburger, A., & Schmitt, M. (2013). Interventions against norm violations: Dispositional determinants of self-reported and real moral courage. *Personality and Social Psychology Bulletin, 39*, 1053–1068. https://doi.org/10.1177/0146167213490032.

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*, 668–686. https://doi.org/10.1111/josi.12085.

Bjørkelo, B., Einarsen, S., Nielsen, M. B., & Matthiesen, S. B. (2011). Silence is golden? Characteristics and experiences of self-reported whistleblowers. *European Journal of Work and Organizational Psychology, 20*, 206–238. https://doi.org/10.1080/13594320903338884.

Bohner, G., Ahlborn, K., & Steiner, R. (2010). How sexy are sexist men? Women's perception of male response profiles in the Ambivalent Sexism Inventory. *Sex Roles, 62*, 568–582. https://doi.org/10.1007/s11199-009-9665-x.

Bosson, J. K., Vandello, J. A., Burnaford, R. M., Weaver, J. R., & Wasti, S. A. (2009). Precarious manhood and displays of physical aggression. *Personality and Social Psychology Bulletin, 35*, 623–634. https://doi.org/10.1177/0146167208331161.

Brandstätter, V., Jonas, K. J., Koletzko, S. H., & Fischer, P. (2016). Self-regulatory processes in the appraisal of moral courage situations. *Social Psychology, 47*, 201–213. https://doi.org/10.1027/1864-9335/a000274.

Branscombe, N. R., Ellemers, N., Spears, R., & Doosje, B. (1999). The context and content of social identity threat. In N. Ellemers, R. Spears, & B. Doosje (Eds.). *Social identity: Context, commitment, content* (pp. 35–58). Oxford, UK: Blackwell Science.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*, 307–324. https://doi.org/10.1037/0033-2909.86.2.307.

Bulman, M. (2018, January 8). H&M apologises following backlash over 'racist' image of child model on website. Retrieved from Independenthttps://www.independent.co.uk/news/uk/home-news/hm-apology-racist-image-website-child-model-backlash-twitter-monkey-jumper-black-a8147641.html.

Cadieux, J., & Chasteen, A. L. (2015). You gay, bro? Social costs faced by male confronters of antigay prejudice. *Psychology of Sexual Orientation and Gender Diversity, 2*, 436–446. https://doi.org/10.1037/sgd0000134.

Cihangir, S., Barreto, M., & Ellemers, N. (2014). Men as allies against sexism: The positive effects of a suggestion of sexism by male (vs. female) sources. *SAGE Open, 4*https://doi.org/10.1177/2158244014539168.

Cottrell, C. A., & Neuberg, S. L. (2005). Different emotional reactions to different groups: A sociofunctional threat-based approach to "prejudice". *Journal of Personality and Social Psychology, 88*, 770–789. https://doi.org/10.1037/0022-3514.88.5.770.

Cowan, G., & Hodge, C. (1996). Judgments of hate speech. The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology, 26*, 355–374. https://doi.org/10.1111/j.1559-1816.1996.tb01854.x.

Cramwinckel, F. M., van Dijk, E., Scheepers, D., & van den Bos, K. (2013). The threat of moral refusers for one's self-concept and the protective function of physical cleansing. *Journal of Experimental Social Psychology, 49*, 1049–1058. https://doi.org/10.1016/j.jesp.2013.07.009.

Czopp, A. M., & Monteith, M. J. (2003). Confronting prejudice (literally): Reactions to confrontations of racial and gender bias. *Personality and Social Psychology Bulletin, 29*, 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*, 784–803. https://doi.org/10.1037/0022-3514.90.5.784.

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*, 112–119. https://doi.org/10.1002/ejsp.855.

Dodd, E. H., Guiliano, T., Boutell, J., & Moran, B. E. (2001). Respected or rejected: Perceptions of women who confront sexist remarks. *Sex Roles, 45*, 567–577. https://doi.org/10.1023/A:1014866915741.

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. *Advances in Experimental Social Psychology, 36*, 4–56. https://doi.org/10.1016/S0065-2601(04)36001-6.

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

Elder, T. J., Sutton, R. M., & Douglas, K. M. (2005). Keeping it to ourselves: Effects of audience size and composition on reactions to criticisms of the ingroup. *Group Processes & Intergroup Relations, 8*, 231–244. https://doi.org/10.1177/1368430205053940.

Eliezer, D., & Major, B. (2012). It's not your fault: The social costs of claiming discrimination on behalf of someone else. *Group Processes & Intergroup Relations, 15*, 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*, 175–191. https://doi.org/10.3758/BF03193146.

Fazio, R. H., & Hilden, L. E. (2001). Emotional reactions to a seemingly prejudiced response: The role of automatically activated racial attitudes and motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 27*, 538–549. https://doi.org/10.1177/0146167201275003.

Fiske, S. T., & Stevens, L. E. (1993). What's so special about sex? Gender stereotyping and discrimination. In S. Oskamp, & M. Costanzo (Vol. Eds.), *Gender issues in contemporary society: Applied social psychology annual. Vol. 6. Gender issues in contemporary society: Applied social psychology annual* (pp. 173–196). Newbury Park, CA: Sage.

Garcia, D. M., Schmitt, M. T., Branscombe, N. R., & Ellemers, N. (2010). Women's reactions to ingroup members who protest discriminatory treatment: The importance of beliefs about inequality and response appropriateness. *European Journal of Social Psychology, 40*, 733–745. https://doi.org/10.1027/1864-9335/a000093.

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*, 653–667. https://doi.org/10.1111/josi.12084.

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist, 56*, 109–118. https://doi.org/10.1037/0003-066X.56.2.109.

Greitemeyer, T., Osswald, S., Fischer, P., & Frey, D. (2007). Civil courage: Implicit theories, related concepts, and measurement. *The Journal of Positive Psychology, 2*, 115–119. https://doi.org/10.1080/17439760701228789.

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. *Psychol. Women Q. 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*, 280–293. https://doi.org/10.1080/15534510.2012.736879.

Haidt, J. (2003). Elevation and the positive psychology of morality. In C. L. M. Keyes, & J. Haidt (Eds.). *Flourishing: Positive psychology and the life well-lived* (pp. 275–289). Washington, DC: American Psychological Association.

Halmburger, A., Baumert, A., & Schmitt, M. (2015). Anger as driving factor of moral courage in comparison with guilt and global mood: A multimethod approach. *European Journal of Social Psychology, 45*, 39–51. https://doi.org/10.1002/ejsp.2071.

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00998.

Hauser, D. J., Paolacci, G., & Chandler, J. (2018). Common concerns with Mturk as a participant pool: Evidence and solutions. Retrieved from https://www.researchgate.net/profile/David_Hauser2/publication/327382170_Common_Concerns_with_MTurk_as_a_Participant_Pool_Evidence_and_Solutions/links/5b8af53b4585151fd14272de/Common-Concerns-with-MTurk-as-a-Participant-Pool-Evidence-and-Solutions.pdf.

Hornsey, M. J., & Esposo, S. (2009). Resistance to group criticism and recommendations for change: Lessons from the intergroup sensitivity effect. *Social and Personality Psychology Compass, 3*, 275–291. https://doi.org/10.1111/j.1751-9004.2009.00178.x.

Hornsey, M. J., & Imani, A. (2004). Criticizing groups from the inside and the outside: An identity perspective on the intergroup sensitivity effect. *Personality and Social Psychology Bulletin, 30*, 365–383. https://doi.org/10.1177/0146167203261295.

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

Jonas, K. J., & Brandstätter, V. (2004). Brennpunkt Zivilcourage: Definitionen, Befunde und Maßnahmen *Focus on moral courage: Definitions, findings, and intervention. Zeitschrift für Sozialpsychologie, 35*, 185–200.

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*, 254–263. https://doi.org/10.1177/0146167201272010.

Kaiser, C. R., & Miller, C. T. (2003). Derogating the victim: The interpersonal consequences of blaming events on discrimination. *Group Processes & Intergroup Relations, 6*, 227–237. https://doi.org/10.1177/1368430203063001.

Kayser, N. D., Greitemeyer, T., Fischer, P., & Frey, D. (2010). Why mood affects help giving, but not moral courage: Comparing two types of prosocial behavior. *European*

Journal of Social Psychology, 40, 1136–1157. https://doi.org/10.1002/ejsp.717.
Landrine, H. (1985). Race × class stereotypes of women. *Sex Roles, 13*, 65–75. https://doi.org/10.1007/BF00287461.
Loughnan, S., Haslam, N., Murnane, T., Vaes, J., Reynolds, C., & Suitner, C. (2010). Objectification leads to depersonalization: The denial of mind and moral concern to objectified others. *European Journal of Social Psychology, 40*, 709–717. https://doi.org/10.1002/ejsp.755.
Major, B., Testa, M., & Bylsma, W. H. (1991). Responses to upward and downward social comparisons: The impact of esteem-relevance and perceived control. In J. Suls, & T. A. Wills (Eds.). *Social comparison: Contemporary theory and research* (pp. 237–260). Hillsdale, NJ: Erlbaum.
Mallett, R. K., & Melchiori, K. J. (2014). Goal preference shapes confrontations of sexism. *Personality and Social Psychology Bulletin, 40*, 646–656. https://doi.org/10.1177/0146167214521468.
Mallett, R. K., & Wagner, D. E. (2011). The unexpectedly positive consequences of confronting sexism. *Journal of Experimental Social Psychology, 47*, 215–220. https://doi.org/10.1016/j.jesp.2010.10.001.
Marques, J. M., & Paez, D. (1994). The 'black sheep effect': Social categorization, rejection of ingroup deviates, and perception of group variability. *European Review of Social Psychology, 5*, 37–68. https://doi.org/10.1080/14792779543000011.
Miller, W. I. (2000). *The mystery of courage.* Cambridge, MS: Harvard University Press.
Minson, J. A., & Monin, B. (2012). Do-gooder derogation: Putting down morally-motivated others to defuse implicit moral reproach. *Social Psychological and Personality Science, 3*, 200–207. https://doi.org/10.1177/1948550611415695.
Monin, B. (2007). Holier than me? Threatening social comparison in the moral domain. *International Review of Social Psychology, 20*, 53–68. Retrieved from http://psych.stanford.edu/~monin/papers/MoninIRSP2007.pdf.
Monin, B., & Jordan, A. H. (2009). The dynamic moral self: A social psychological perspective. In D. Narvaez, & D. K. Lapsley (Eds.). *Personality, identity, and character: Explorations in moral psychology* (pp. 341–354). New York, NY: Cambridge University Press.
Monin, B., Sawyer, P., & Marquez, M. (2008). The rejection of moral rebels: Resenting those who do the right thing. *Journal of Personality and Social Psychology, 95*, 76–93. https://doi.org/10.1037/0022–3514.95.1.76.
Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review, 110*, 472.
Nail, P. R., Harton, H. C., & Decker, B. P. (2003). Political orientation and modern versus aversive racism: Tests of Dovidio and Gaertner's (1998) integrated model. *Journal of Personality and Social Psychology, 84*, 754–770. https://doi.org/10.1037/0022-3514.84.4.754.
O'Connor, K., & Monin, B. (2016). When principled deviance becomes moral threat: Testing alternative mechanisms for the rejection of moral rebels. *Group Processes & Intergroup Relations, 19*, 676–693. https://doi.org/10.1177/1368430216638538.
Phelan, J. E., Moss-Racusin, C. A., & Rudman, L. A. (2008). Competent yet out in the cold: Shifting criteria for hiring reflect backlash toward agentic women. *Psychology of Women Quarterly, 32*, 406–413. https://doi.org/10.1111/j.1471-6402.2008.00454.x.
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*, 8–16. https://doi.org/10.1080/01973530903539754.
Rodin, M. J., Price, J. M., Bryson, J. B., & Sanchez, F. J. (1990). Asymmetry in prejudice attribution. *Journal of Experimental Social Psychology, 26*, 481–504. https://doi.org/10.1016/0022-1031(90)90052-N.
Rudman, L. A., Mescher, K., & Moss-Racusin, C. A. (2013). Reactions to gender egalitarian men: Perceived feminization due to stigma-by-association. *Group Processes & Intergroup Relations, 16*, 572–599. https://doi.org/10.1177/1368430212461160.
Schnall, S., Roper, J., & Fessler, D. M. (2010). Elevation leads to altruistic behavior. *Psychological Science, 21*, 315–320. https://doi.org/10.1177/0956797609359882.
Sekerka, L. E., & Bagozzi, R. P. (2007). Moral courage in the workplace: Moving to and from the desire and decision to act. *Business Ethics: A European Review, 16*, 132–149. https://doi.org/10.1111/j.1467-8608.2007.00484.x.
Shelton, J. N., & Stewart, R. E. (2004). Confront perpetrators of prejudice: The inhibitory effects of social costs. *Psychology of Women Quarterly, 28*, 215–223. https://doi.org/10.1111/j.1471-6402.2004.00138.x.
Skitka, L. J. (2011). Moral convictions and moral courage: Common denominators of good and evil. In M. Mikulincer, & P. R. Shaver (Eds.). *The social psychology of morality: Exploring the causes of good and evil* (pp. 349–365). Washington, D.C: American Psychological Association.
Swim, J. K., & Hyers, L. L. (1999). Excuse me—What did you say?! Women's public and private responses to sexist remarks. *Journal of Experimental Social Psychology, 35*, 68–88. https://doi.org/10.1006/jesp.1998.1370.
Walker, L. J. (1999). The perceived personality of moral exemplars. *Journal of Moral Education, 28*, 145–162. https://doi.org/10.1080/030572499103188.
Wood, J. V. (1989). Theory and research concerning social comparison of personal attributes. *Psychological Bulletin, 106*, 231–248. https://doi.org/10.1037/0033-2909.106.2.231.
Woodzicka, J. A., & LaFrance, M. (2001). Real versus imagined gender harassment. *Journal of Social Issues, 57*, 15–30. https://doi.org/10.1111/0022-4537.00199.