

Mass Communication and Society

ISSN: 1520-5436 (Print) 1532-7825 (Online) Journal homepage: www.tandfonline.com/journals/hmcs20

White Apathy and Allyship in Uncivil Racial Social Media Comments

Jessica Roden & Muniba Saleem

To cite this article: Jessica Roden & Muniba Saleem (2022) White Apathy and Allyship in Uncivil Racial Social Media Comments, Mass Communication and Society, 25:3, 383-406, DOI: 10.1080/15205436.2021.1955933

To link to this article: https://doi.org/10.1080/15205436.2021.1955933

View supplementary material

Published online: 06 Aug 2021.

Submit your article to this journal

Article views: 1212

View related articles

View Crossmark data

Citing articles: 6 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hmcs20

MASS COMMUNICATION AND SOCIETY
2022, VOL. 25, NO. 3, 383–406
https://doi.org/10.1080/15205436.2021.1955933

Routledge
Taylor & Francis Group

Check for updates

White Apathy and Allyship in Uncivil Racial Social Media Comments

Jessica Roden [a] and Muniba Saleem [b,*]

[a]Communication and Media, University of Michigan, Ann Arbor, MI, USA; [b]Communication Studies, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Given the pervasiveness of racist online comments, it is important to determine what factors motivate White audiences to take actions for racial justice. We utilize and expand on the defensive effect, which proposes that uncivil online comments can mobilize action, by focusing on social identities and group-based emotions. By manipulating the race of a person tweeting in support of racial justice protests and the civility of the racist response, we find that uncivil replies lead to increased feelings of group-based anger, which is associated with willingness to respond to the comment online and take political action for racial equality. These findings highlight the mobilizing influence that anger, as opposed to other emotions, can have on White bystanders witnessing racist incivility within social media environments.

Online discrimination and harassment are rampant, with 41% of Americans having been subjected to some sort of online harassment and one in four Black individuals experiencing online harassment due to their race (Duggan, 2017). However, only half of those who ever witnessed online harassment took any action (Duggan, 2017). While bystanders might think that they would feel upset and stand up for the victim or address the perpetrator when witnessing racism, in reality they feel little distress and are unlikely to intervene (Kawakami et al., 2009). Because racist online comments are common and bystander intervention less so, it is important to determine the mechanisms that do inspire action, both on and offline, to understand how the harmful effects of these comments might be mitigated.

CONTACT Jessica Roden jeroden@umich.edu Communication and Media, University of Michigan, North Quad, Room 5370, 105 South State Street, Ann Arbor, MI 48109, USA.
*Present affiliation: Department of Communication, University of California, Santa Barbara, Santa Barbara, CA, USA.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/15205436.2021.1955933

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

© 2021 Mass Communication & Society Division of the Association for Education in Journalism and Mass Communication

Prior literature has demonstrated that, in personally relevant situations, these hurtful or uncivil online comments evoke negative emotions that can motivate willingness to take action (Borah, 2014; Lu & Myrick, 2016; Masullo Chen, 2017; Masullo Chen & Lu, 2017). This motivation to defend one's beliefs when attacked is known as the defensive effect (Masullo Chen, 2017). Witnessing a racist conversation, though, is not as much personally relevant as it is relevant to an individual's social identity, whether that be through a common racial identity, political ideology, or other salient aspect of the remarks. While uncivil intergroup online comments have been examined (J. W. Kim, 2018; Rains et al., 2017), the defensive effect, to our knowledge, has not been extended to an explicitly intergroup setting, in which threatening intergroup conflict heightens social identity salience and provides opportunities to react as a member of a social group (Spears, 2011).

This experiment investigates this context by having White participants witness a racist online argument. White people, as those with the most control over American institutions, have the most opportunity and responsibility to enact racial equality (DiAngelo, 2018). It is therefore necessary to determine how this racial group will react to witnessing racist social media comments. If salient social identity cues, such as race, are present that link the bystander to the perpetrator of the racist comment, the bystander is likely to experience the interaction as someone of that salient social identity group (Mackie et al., 2016). In terms of the defensive effect, we expect this heightened social identity connection to be manifested in negative group-based emotions, which are emotions experienced on behalf of a group member, and racial justice action intentions. To test the mechanisms mobilizing White people upon seeing racist online incivility, we used an experimental paradigm that manipulates both the race of the person who tweets a message in support of athletes kneeling to protest police brutality (White or Black) and the civility of the reply from a White person who makes a racist comment in disagreement with this stance (neutral or uncivil).

This study contributes theoretically by building on understanding of the defensive effect. Instead of treating all online conversations as only interpersonal, this work determines if and how this effect translates to an intergroup situation by highlighting the role of social identities and group-based emotions. By combining literature on social identity theory with research about the defensive effect, we are able to better understand the effects of racist online incivility and the specific group-based emotions that help mobilize action for racial justice among White people. Further, by examining individual group-based emotions and their distinct effects, we add nuance to the mechanisms involved in reactions to racist online incivility. Practically, this study will be useful in determining the conditions

needed to mobilize White individuals to take anti-racism action in social media settings and beyond.

Online incivility and the defensive effect

Online incivility is any remark that conveys an unnecessarily disrespectful tone (Coe et al., 2014), including either insulting language, name-calling, profanity, stereotypes, and other terms related to hate speech (Masullo Chen, 2017). This is a pervasive problem; in fact, one in five online comments has been found to exhibit at least some attribute of uncivil speech (Coe et al., 2014). These comments can have serious negative outcomes, including greater close-mindedness, attitude certainty, and prejudice (Borah, 2014; Hwang et al., 2018; Soral et al., 2018), which limit opportunities for democratic discourse.

While uncivil comments from politicians tend to increase cynicism and demobilize constituents (Mutz, 2015; Otto et al., 2020), in certain specific situations, uncivil online comments from ordinary non-elites can actually have a mobilizing effect. Masullo Chen (2017) labels this the defensive effect, since incivility can cause people to want to defend their beliefs from being attacked. The message must be (1) a disagreement, (2) that contains incivility, (3) about a polarizing issue with clear sides that the person cares deeply about, and (4) in a way that feels personal or challenges the person's sense of self (Masullo Chen, 2017). If this type of message evokes a strong negative affect in someone who deeply disagrees with it, this person is likely to want to defend their point of view by speaking out or getting politically involved with the issue at focus.

Several studies have demonstrated the link between these kinds of uncivil comments and intentions to participate politically on a wide range of issues including abortion, gay rights, immigration policy, gun control, military intervention with ISIS, removal of Confederate flags, and a presidential campaign (Borah, 2014; Lu & Myrick, 2016; Masullo Chen, 2017; Masullo Chen & Lu, 2017). Several of these studies found this association to exist only indirectly through negative emotions, whereas a lack of these emotions produced a sort of silencing effect, inhibiting action intentions (Lu & Myrick, 2016; Masullo Chen, 2017). Regardless of the topic, when individuals felt personally challenged by uncivil disagreement and experienced negative emotions, they felt more inclined to take action for the issue they cared about.

Intergroup dynamics involved in online racial dialogue

Would a message be able to trigger similar defensiveness when someone feels challenged, not personally, but due to a shared social identity? The processes

related to social identity connections are outlined in social identity theory, which suggests that an identity can originate from group membership when it has important emotional and value significance, leading to the creation of an "us" versus "them" (Tajfel & Turner, 1979). This has important consequences because an individual is more likely to align with the ingroup's attitudes and norms and to reject those of the outgroup in order to maintain positive distinctiveness (Spears, 2011). These social identities are made salient in intergroup contexts, in which group categories are evident and magnify intergroup differences above interpersonal ones (Spears, 2011).

Subtle cues like background music or photos are often all that is needed to prime a social identity in these contexts (Seger et al., 2009). While there might be several relevant social identities in a given situation, race is one that is both chronically and situationally accessible, meaning that it is salient both because it is important to one's self-concept and very perceptually evident (Hogg & Reid, 2006). White individuals categorize based on race more quickly than they would like to admit (Norton et al., 2006), and race has been found to be a more salient factor in categorizing facial images than gender and age (Montepare & Opeyo, 2002). Especially when a conversation is clearly connected to racism, which it is in the case of athletes kneeling during the national anthem to protest police brutality (Boycoff & Carrington, 2020), there are several cues available to prime racial social identity.

The racism produced from the threats these cues create contributes to many harmful intergroup behaviors. In the presence of other bystanders, White people feel less compelled to help out a Black victim compared to a White one (Gaertner & Dovidio, 1977; Saucier et al., 2005). In fact, there is no evidence that people who witness racist incidents experience behavioral, cognitive, or physiological reactions of being upset (Karmali et al., 2017). Further, the majority of bystanders do not confront the individual who makes a racist comment, and while feeling strong negative emotions encourages confrontation, especially for women, it does not guarantee action (Dickter & Newton, 2013).

Similar effects also occur in online contexts. People report seeing hate speech on social media at a greater rate than during face-to-face communication (Barnidge et al., 2019). Online comments about the tragic extrajudicial killings of Black Americans, for example, contain more incivility than they do thoughtful deliberation on the subject (Masullo Chen et al., 2020). Witnessing these comments might just lead to apathy though, because maliciousness in online environments evokes less outrage than it does in face-to-face contexts (Puryear & Vandello, 2019). It is therefore important to examine the emotions that are produced in these settings to see if any can lead to action against racism.

Discrete group-based emotions

The appraisal theories of emotions recognize that emotions are derived from evaluations of events, which can differ between individuals (Scherer et al., 2001). In the context of the defensive effect, multiple people could read the same uncivil comment, but only those who hold strong beliefs about the issue and disagree with the author would evaluate the comment as a threat to their beliefs and respond with negative emotions. Intergroup appraisals occur when a certain social identity is made salient, group members interpret a situation as a member of that identity group, and they think about implications as a group member (Mackie et al., 2016). While salience can be activated by explicit mentions of a social identity, it can also occur through the priming of more subtle group cues (Seger et al., 2009).

According to Intergroup Emotions Theory, these intergroup appraisals lead to intergroup or group-based emotions, which are emotions experienced on behalf of a group during group-relevant events (Mackie et al., 2000). These are distinct from individual emotions because they can be experienced in response to situations that affect ingroup members but not oneself, such as feeling anger when an ingroup member is harmed by another person, even when the harm does not personally affect the observer (Gordijn et al., 2001; Smith et al., 2007). This distinction is important because group-based emotional reactions connect emotions to social identity categorizations and experiences as an ingroup member, contexts which are not captured by individual emotions (Mackie et al., 2016).

Since the same situation can evoke multiple different group-based emotions simultaneously (Mackie et al., 2016), the group-based emotions examined in this experiment are anger, guilt, and shame. These emotions can lead to distinct actions, including ones that are key for anti-racism, like intervening with the racist incident online and engaging politically to combat racism at a larger, structural scale. With bystander anti-racism reducing the harm to victims and potentially shifting social norms away from intolerance (Nelson et al., 2011), and political anti-racism action influencing electoral and policy choices as well as creating a sense of community (Norris, 2007), the two actions together have the opportunity to create meaningful change.

Anger

Anger has been studied extensively in the collective action literature as one of the essential antecedents of action. Indeed, the social identity model of collective action (SIMCA) considers anger evoked by perceived injustice to be a key mechanism in motivating collective actions (Van Zomeren et al., 2008).

According to the appraisal theories of emotions, anger is especially likely to be experienced when a situation is appraised as disadvantageous and in violation of social norms (Kuppens et al., 2007). When such injustice or violation is perceived to occur to the ingroup, individuals feel angry at this mistreatment and are likely to seek actions that change or challenge their disadvantageous position. Though much existing work examines these processes among disadvantaged group members (Saleem et al., 2021), research suggests advantaged group members also display similar processes when supporting disadvantaged groups (Leach et al., 2006; Van Zomeren et al., 2011).

Anger has also been found to be a mechanism for action in relation to uncivil comments. Online messages that are uncivil or racist, understandably, produce feelings of anger (Lee-Won et al., 2017). Hateful comments can include norm violations such as insults and defamations, calls for violence, and political agitation (Wilhelm et al., 2020), appraisals of which are known to induce anger (Kuppens et al., 2007). When cross-cutting and uncivil comments evoke anger, this anger then leads to intentions to engage in low-cost political activities like sharing a news article or engaging in political discussions on social media (Lu & Myrick, 2016). This study found that anger did not, however, lead to more costly participation intentions like donating money or volunteering, most likely because anger triggers quick responses that do not require much contemplation (Lerner & Keltner, 2001; Scott, 1980) – the opposite of the dedication needed to sustain commitment for these high-cost activities.

Shame and guilt

While colloquially shame and guilt might be used as synonyms, prior research has found them to be distinct emotions that have different antecedents and outcomes (Lickel et al., 2005). When someone has a shared identity with another who committed an act of wrongdoing, if it seems like it could hurt one's self-image, the person is likely to feel shame and take distancing actions. Group-based shame is related to the black sheep effect, whereby an ingroup member with negative qualities gets derogated more than an outgroup member with the same qualities (Marques et al., 1988). This is often done in an effort to distance that person from the group while retaining a positive social identity (Marques & Paez, 1994). For example, following the September 11 attacks, Americans who witnessed discrimination by fellow Americans toward Middle Easterners felt group-based shame and wanted to distance themselves from their fellow Americans (Johns et al., 2005).

On the other hand, if an individual has perceptions of interdependence, or high levels of interaction and communication, with the wrongdoer, the person might feel responsible for not stopping the event. This, in turn, will produce group-based guilt and reparative actions to get involved and fix the problem

(Mallett et al., 2008). Because of this perceived interdependence, it seems as if group-based guilt would not be evoked during an uncivil encounter with an internet stranger. However, research has found group-based guilt to be provoked when memories of historical wrongdoings are brought up, even though the individuals who feel this guilt could not have done anything to stop the past (Brown et al., 2008; Doosje et al., 1998). In the context of the United States, the collective guilt that White people might feel about historical and present mistreatment and abuse of Black Americans could be seen as *White guilt*. While White guilt is only present in a small portion of White people, feeling this guilt about past and present inequality toward Black Americans has been found to predict affirmative action support to try to remedy those racist actions and the guilt they cause (Swim & Miller, 1999).

Discrete emotions

In the context of this experiment, in which an ingroup member makes a racist comment, these three group-based emotions are likely to be evoked and lead to different action intentions. The defensive effect, as Masullo Chen (2017) specifies it, however, does not predict how incivility will impact discrete emotions, but just how it will influence general negative affect. It is theorized as relying on primal emotional responses, not higher-level cognitive processes, making it impossible to disentangle one emotion from another (Masullo Chen, 2017). For example, in one study, negative affect is operationalized by asking if participants felt "upset," "guilty," "jittery," "ashamed," "depressed," "angry," "irritable," or "annoyed" (Masullo Chen & Lu, 2017). However, this seems limiting, as at ignores how events are uniquely appraised and how distinct emotions can lead to different outcomes (Nabi, 2010; Scherer et al., 2001). This is demonstrated in one study that found that both cross-cutting arguments and incivility on Facebook created anxiety, which predicted engagement in high-cost activities like donating money or volunteering, and anger, which predicted low-cost activities like talking to others (Lu & Myrick, 2016). These nuances would be missed by treating all negative emotions as one. By highlighting discrete group-based emotions, this study examines these distinctions with respect to racist incivility.

Hypotheses

The literature clearly points to personally relevant uncivil attacks evoking strong negative emotions that prompt action (Lu & Myrick, 2016; Masullo Chen, 2017). We hypothesize that the same will occur for racist incivility that is relevant not to the personal, but to the social sense of self. Further, prior research has found that an ambiguous negative situation between a Black and White person is labeled as a conflict more often than the

same situation featuring two White individuals (Karmali et al., 2019). Yet, people often react apathetically upon witnessing racism (Karmali et al., 2017). We therefore investigate how disagreeing comments directed at a White versus a Black person alter negative emotions, and how the degree of incivility of the comments might interact with this race manipulation to influence emotional responses.

H1: Uncivil replies, compared to neutral replies, will lead to more group-based anger (H1a), guilt (H1b), and shame (H1c).

RQ1: What effect will replies targeted at a Black person compared to a White person have on group-based anger (RQ1a), guilt (RQ1b), and shame (RQ1c)?

RQ2: How will the civility manipulation interact with the race manipulation to influence group-based anger (RQ2a), guilt (RQ2b), and shame (RQ2c)?

While collectively, greater negative emotions should lead to increased action, this study seeks to isolate these emotions to determine if each will lead to distinct action intentions. It is expected that group-based anger will lead to political action and comment response intentions because it is a well-known predictor of collective action and is useful in propelling quick, low-cost actions (Lu & Myrick, 2016; Van Zomeren et al., 2008). Group-based guilt should also increase political action intentions in order to help repair the racist damage done, but guilt is less likely to lead to comment response intentions since it has less of the fiery motivation needed to comment on the conversation (Mallett et al., 2008). Finally, group-based shame, as an emotion that leads to distancing actions (Johns et al., 2005), should not predict either action intention.

H2: The uncivil condition will increase political action intentions via group-based anger (H2a) and guilt (H2b), and this indirect relationship will be greater than that for group-based shame (H2c).

H3: The uncivil condition will indirectly increase comment response intentions via group-based anger (H3a), and this indirect relationship will be greater than that for group-based guilt (H3b) or shame (H3c).

RQ3: What indirect effects will the race manipulation have on political action intentions (RQ3a) and comment response intentions (RQ3b) via these negative emotions?

Method

Participants

The Institutional Review Board at the University of Michigan gave this study an exempt approval determination on August 31, 2018. We recruited

380 students from the Communication Studies and Psychology subject pools of the University of Michigan in the United States. Because the focus of this research is on the responses of White individuals to racist incivility, analyses were done on White participants. Data from 57 non-white students were deleted from the dataset, along with two participants who did not indicate their gender, one who reported the age of 17, despite prescreen requirements that limited the age to 18, and three with missing data. On average, the remaining 317 participants were 63.7% female with a mean age of 18.83 (SD = 0.94) and identifying as slightly liberal on a scale of 1 = "strongly conservative" to 7 = "strongly liberal" (M = 4.87, SD = 1.56).

Procedure

After providing informed consent, participants first rated several issues to assess which ones were most important to them. They were told that they would be shown two Twitter conversations based on their answers to these questions, but regardless all participants were first shown a conversation about homelessness, to distract from the true intention of the study, and then racial justice. These conversations were photo-edited to look like real tweets by adding in text of a tweet and a stock image for a profile photo. We made race a very salient social identity by focusing the conversation on racial justice protests and cueing racial identities with names and profile photos. The conversation was about "campus plans to support NFL players who kneeled to protest police brutality," with the first person in support of this plan and the second person opposed. This topic was selected because it is both very racially and politically divisive (Casteel, 2017), increasing the possibility that it will lead to intergroup conflict. Participants were randomly assigned to one of four conditions, with the initial tweet either from a Black or White person, and the responding tweet (always from a White person) either neutral or uncivil. After exposure to one of these conversations about a racial justice issue, participants answered a series of questions assessing outcomes of interest.

To further increase the likelihood of ingroup identification from shared salient features (Tajfel & Turner, 1979), the stimuli also indicated shared university affiliation and gender identity. The conversation was introduced as a response to campus plans and was a reply to the well-known campus newspaper Twitter account, signaling that the accounts discussing this issue were likely affiliated with the participants' university. The gender of the individuals portrayed in the conversation was matched to the participant's self-identified gender by changing the gender presented in the profile photos and first names of the two people in the Twitter conversation. While some participants might have identified more with the person who

supported the campus plans due to shared opinions on this topic, the emphasis on race and common racial identity along with shared university affiliation and gender identity makes it very likely that an ingroup social identity was activated to some degree with the person opposing the campus plans.

Stimuli

The first tweet, which was an actual tweet taken from Twitter, was supportive of the NFL players kneeling and stayed the same across conditions. The uncivil and neutral replies were also real tweets found on Twitter, although slightly modified for length and grammar. The uncivil tweets contained either name-calling, profanity, or capital letters to indicate a raised tone, all of which are common indicators of incivility (Coe et al., 2014; Papacharissi, 2004) and align with the types of uncivil stimuli used in prior work (Masullo Chen & Lu, 2017). Name-calling was incorporated in all versions, as it is the most prevalent form of incivility (Coe et al., 2014). The neutral tweets were about the topic of NFL players kneeling during the national anthem and were selected for their disagreement with the stance these players were taking without being uncivil about it. In order to increase generalizability, three uncivil replies and three neutral replies were selected so one could be randomly shown to participants within each condition. Similar to Masullo Chen and Lu (2017) 20 comments were evaluated in a pretest using a positive face threat scale, with higher numbers indicating greater incivility. The three comments with the highest and lowest means as well as the most reliable index were chosen as stimuli for this experiment (Masullo Chen & Lu, 2017). Because the uncivil comments on their own were not directly related to the NFL protests, the experimental stimuli were created by adding one of the three uncivil comments to the beginning of one of the three neutral comments to create the uncivil conditions (Lu & Myrick, 2016). The various comments were rigorously pretested to ensure that they were appropriately neutral or uncivil (see Appendix A of supplementary analyses). See Appendix B for the comments used.

Along with manipulating the civility of the replies, the race of the person making the first tweet was also manipulated. Racial identities were cued by changing the photo and name of the person whom the tweet was from. Similar to Lane et al. (2019), names were selected based on prior work about stereotypical White and Black names (Greenwald et al., 1998) and Census information about the most common names for different racial groups (Comenetz, 2016). For the male participant condition, the names were Brad Krueger (White) or Tyrone Jefferson (Black) receiving a reply

from Justin Miller (White). For female participants, the names were Amanda Krueger (White) or Tanisha Jefferson (Black) receiving a reply from Courtney Miller (White). The photos were also thoroughly pretested to ensure that only the race manipulation was influencing the outcomes and not other characteristics present in the profile photos of the Twitter users (see Appendix A of supplementary analyses). Both the highest and lowest rated photos for each condition were selected so that participants randomly assigned to each condition would then be randomly assigned to see either the low or high-rated photo for the person initiating the Twitter conversation.[1]

Measures

Table 1 displays descriptive statistics and correlations of key measures. See Appendix C of supplementary analyses for a list of items included in each measure.

Demographic measures

Prior to the experimental manipulation, participants were asked about their gender, in order to sort them into the correct gender-matched condition, and age. At the end, they were asked for their race and political orientation.

Table 1. Descriptive statistics and bivariate correlations for measures (N = 317).

Measure	1	2	3	4	5
1. Anger	–				
2. Guilt	0.48***	–			
3. Shame	0.65***	0.76***	–		
4. Political Action Intentions	0.33***	0.14*	0.23***	–	
5. Comment Response Intentions	0.36***	0.18**	0.23***	0.55***	–
Range	1–5	1–5	1–5	1–7	1–5
Means	2.47	1.40	1.79	3.37	2.25
Standard Deviations	0.94	0.62	0.84	1.50	0.78
Cronbach alphas	0.90	0.77	0.84	0.93	0.80

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

[1]To ensure that the three different comments and two different photos within each condition were not significantly different on any of the outcome variables, a series of ANOVAs and t-tests were computed. No significant differences were found between the three neutral or the three uncivil comments within each condition (all F-values < 2.29, all p-values > .10). There were also no significant differences between the photos for any of the variables (all t-values < 1.95, all p-values > .05). All subsequent analyses were done collapsing across the three neutral comments and across the three uncivil comments, as well as collapsing across the high and low-rated photos.

Prior importance of racial justice

Prior to the experimental manipulation, participants evaluated personal importance of various social issues, adapted from Y. M. Kim (2009), one of them being racial justice, on a scale of 1 (*not at all important*) to 5 (*extremely important*).

Group-based emotions

While some studies measure group-based emotions by explicitly invoking social identity connections (Brown et al., 2008), others ask how the participants would feel in response to a particular situation that does not personally involve the participant (Shepherd et al., 2013; Van Zomeren et al., 2011), which is found to elicit similar responses as the explicit measures when following a group cue (Seger et al., 2009). This study followed the inconspicuous route and asked participants, after the experimental manipulation, to rate to what extent reading those comments made them feel a variety of different emotions on a scale from 1 (*not at all*) to 5 (*extremely*). Anger was measured with items adapted from Shepherd et al. (2013), consisting of "angry," "annoyed," "outraged," and "furious" (α = 0.90). Guilt and shame were evaluated with items adapted from Lickel et al. (2005). Guilt was assessed via the three emotions "guilty," "regret," and "remorse" (α = 0.77), and shame was assessed with "ashamed," "embarrassed," "disgraced," and "humiliated" (α = 0.84).

Political action intentions

Participants were asked about their willingness to take 11 political actions relating to racial justice such as "volunteer time at an organization that supports my views on this issue" and "join a Facebook group for a student organization that supports my views on this issue." These items, adapted from both Masullo Chen and Lu (2017) and Lu and Myrick (2016), modified slightly to be more specific to a university population, asked about willingness to take any of these actions in the next 12 months on a scale from 1 (*extremely unlikely*) to 7 (*extremely likely*), (α = 0.93).

Comment response intentions

Participants answered how likely they would be to engage in a series of online actions in response to these Twitter comments, such as "criticize one of the people by publicly posting a response to their comment" and "send a direct message to one of the people agreeing with their comment." We were less concerned about the target of the criticism or agreement, and more interested in seeing if any online actions would be taken at all. Nine items ranging from 1 (*extremely unlikely*) to 5 (*extremely likely*) were averaged into an index (α = 0.80).

Manipulation check

At the end of the survey, participants were reminded of the second comment in the Twitter conversation they saw that was either neutral or uncivil. They were asked to evaluate this comment on a scale from 1 (*extremely uncivil*) to 7 (*extremely civil*).

Results

Preliminary analysis

Participants in the White neutral (n = 78), White uncivil (n = 82), Black neutral (n = 83), and Black uncivil (n = 74) conditions did not significantly differ from each other in terms of political orientation, age, or prior importance of racial justice. The manipulation check showed that the civility manipulation was successful, with the three neutral comments rated as more civil than the three uncivil comments (see Appendix A for details).

Main results

A MANOVA with political action intentions, comment response intentions, group-based anger, group-based guilt, and group-based shame as outcomes and civility, race, and prior importance of racial justice as predictors revealed a significant difference for civility, $F(5, 308) = 2.75$, $p = .02$, and prior racial justice importance, $F(5, 308) = 15.35$, $p < .001$. To determine how the uncivil condition would affect each outcome, follow-up univariate analyses were performed. See Appendix A for a robustness check using bootstrapped analyses that produced the same substantive findings.

Group-based emotions

ANCOVAs predicting each of the three group-based emotions (anger, guilt, and shame) were computed with race of initial tweet, civility of responding tweet, and interactions between the two as well as prior importance of racial justice as a control included. Figure 1 shows means and confidence intervals for each condition and each group-based emotion.

Both the civility manipulation, $F(1, 312) = 10.32$, $p = .001$, and the prior importance of racial justice, $F(1, 312) = 13.10$, $B = 0.21$, $p < .001$, were significant predictors of participants' anger. Consistent with H1a, those who were exposed to the uncivil stimuli (M = 2.71, SD = 0.94, 95% CI = [2.56, 2.86]) became significantly more angry than those who were exposed to the neutral stimuli (M = 2.23, SD = 0.87, 95% CI = [2.09, 2.36]). Contrary to H1b, there were no significant effects of civility predicting feelings of guilt. As for feelings of shame, consistent with H1c, the civility manipulation was

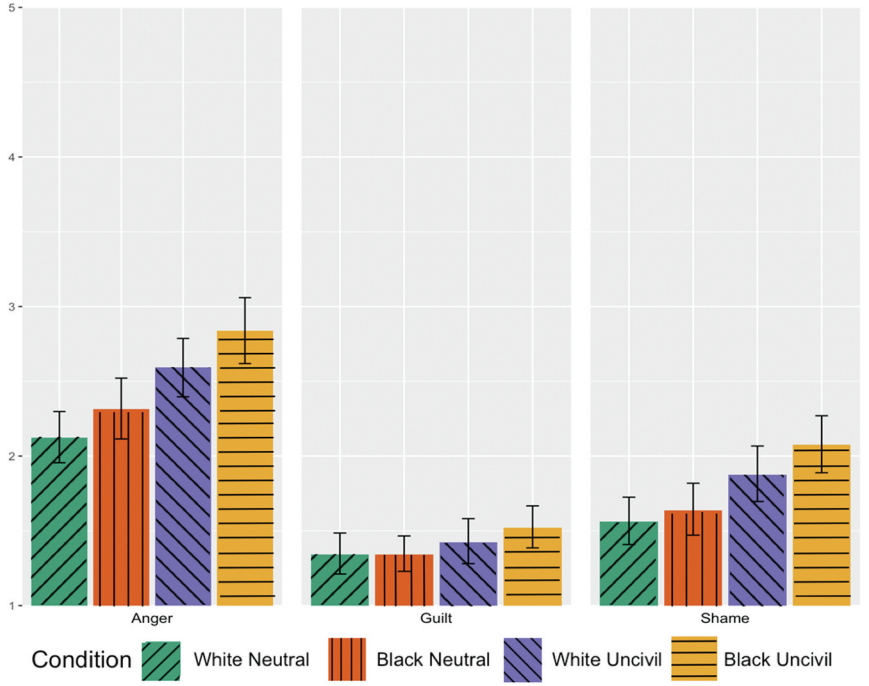


Figure 1. The effects of civility and race conditions on group-based anger, guilt, and shame.

Note. This figure is shown with 95% bootstrapped confidence intervals.

a significant predictor, $F(1, 312) = 5.66$, $p = .02$, with those seeing the uncivil stimuli ($M = 1.97$, $SD = 0.86$, 95% CI = [1.84, 2.11]) experiencing greater group-based shame than those seeing the neutral stimuli ($M = 1.60$, $SD = 0.77$, 95% CI = [1.48, 1.72]). Answering RQ1 and RQ2, neither the race manipulation nor the interaction of race and civility predicted anger, guilt, or shame.

Political action intentions and comment response intentions

To test H2 and H3, looking for the indirect effects of civility on political action intentions and comment response intentions through the various emotions, we used the R package *Mediation*. This uses a 10,000 simulation non-parametric bootstrapping technique to generate point estimates and 95% confidence intervals (Tingley et al., 2014). Parallel mediation models were created with the predictors of race, civility, and prior importance of racial justice, and the potential mediators of anger, guilt, and shame, for each comment response and political action intentions.

Consistent with H2a, and as seen in Figure 2, when compared to the neutral conditions, the uncivil conditions led to increased political action

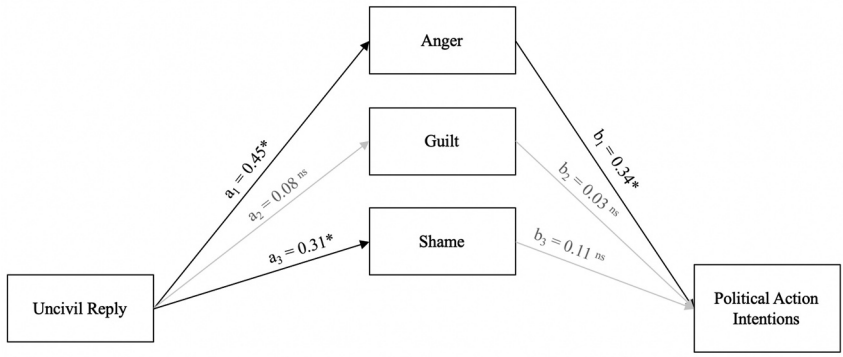


Figure 2. Parallel mediation model showing the uncivil reply (compared to the neutral reply) leading to political action intentions via group-based anger.

Note. *$p < .05$, ns = non-significant. This was performed on 10,000 bootstrap samples, controlling for the race manipulation, the interaction between race and civility, and prior importance of racial justice.

intentions indirectly through increased group-based anger (Indirect Effect = 0.17, 95% CI = [0.06, 0.32], $p = .002$). Neither guilt (Indirect Effect = 0.00, 95% CI = [−0.06, 0.06]) nor shame (Indirect Effect = 0.04, 95% CI = [−0.08, 0.19]) yielded a significant mediating effect in this model. Follow-up analyses computing a bootstrapped mediation effect difference revealed that whereas the indirect effect of anger and guilt may be different from each other, 95% CI = [0.04, 0.33], the difference between anger and shame is not statistically significant, 95% CI = [−0.06, 0.32]. Indeed, whereas the upper confidence interval boundary of guilt matches the lower confidence interval boundary of anger, suggesting a differential effect of these two emotions, the confidence interval of shame has considerable overlap with the confidence interval of anger, suggesting these two emotions are likely similar in their effects in this model. These analyses refute H2b and H2c.

Supporting H3a, the uncivil conditions compared to the neutral conditions also increased comment response intentions indirectly through increased group-based anger, as seen in Figure 3 (Indirect Effect = 0.13, 95% CI = [0.06, 0.22], $p < .001$). Neither guilt (Indirect Effect = 0.01, 95% CI = [−0.02, 0.04]) nor shame (Indirect Effect = −0.01, 95% CI = [−0.09, 0.05]) were significant mediators in this model. Follow-up analyses computing a bootstrapped mediation effect difference revealed that the indirect effect of anger is significantly different than that of guilt, 95% CI = [0.05, 0.21], and shame, 95% CI = [0.05, 0.26]. Indeed, the upper confidence interval boundary of both guilt and shame did not overlap with the lower confidence interval boundary of anger, providing further support of the

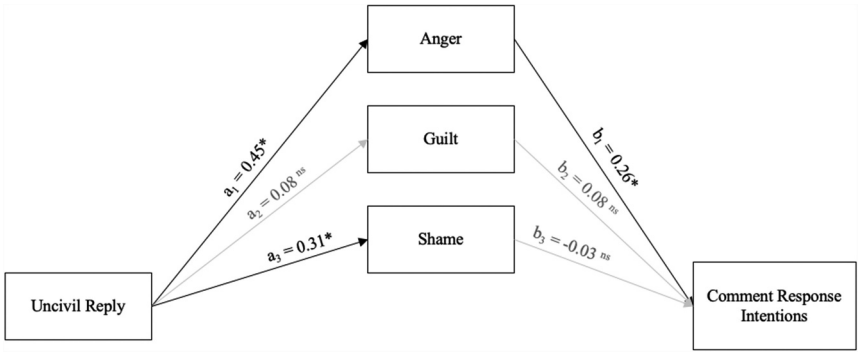


Figure 3. Parallel mediation model showing the uncivil reply (compared to the neutral reply) leading to comment response intentions via group-based anger.

Note. *$p < .05$, ns = non-significant. This was performed on 10,000 bootstrap samples, controlling for the race manipulation, the interaction between race and civility, and prior importance of racial justice.

differential effect of anger compared to guilt and shame. These results support H3b and H3c.

Answering RQ3a and RQ3b, because neither the race condition nor the interaction between race and civility resulted in any negative emotions, it was not tested for indirectly influencing either action intention.

Discussion

Prior research has demonstrated how incivility in personally relevant political contexts can trigger negative emotions or aggressive intentions which can lead to political mobilization (Borah, 2014; Lu & Myrick, 2016; Masullo Chen, 2017; Masullo Chen & Lu, 2017). Witnessing a racist social media interaction, because of social identity cues, leads to experiencing the event as a member of a salient social identity group. We build upon theoretical understanding of the defensive effect by expanding it to an intergroup context and demonstrating the utility of social identities and specific group-based emotions for mobilizing racial justice action intentions.

We found that White participants viewing an uncivil conversation experienced more group-based anger and shame, but not guilt, than those viewing a neutral conversation. The White participants were more outraged and ashamed at this uncivil attack perpetrated by an ingroup member, likely due to a sense of shared social identity that the participants wanted to distance themselves from in order to avoid the repercussions of being associated with this ingroup member. Note however that even with these condition differences, the mean on the 1–5 emotion scale was always below

"moderately" feeling group-based anger and closer to feeling only "a little" group-based shame, meaning the White participants in this study only experienced minimal negative emotions about the racist exchange, even when it was uncivil.

The minimal group-based guilt also experienced by participants in this study along with its null results are consistent with low levels of White guilt found in prior work (Swim & Miller, 1999). Perhaps the stimuli did not capture the interdependence between the participant and commenter enough to make the connection between racism and the self salient in order to produce group-based guilt. It is also likely, though, that most incidents of racism perpetrated by an ingroup member are unlikely to make White people feel group-based guilt, especially if they use the false good/bad dichotomy to separate out that racist user as a bad person and themselves as good non-racists (DiAngelo, 2018). Using this binary could comfort participants into thinking they are no longer members of the same salient social identity, thereby weakening perceptions of interdependence and limiting feelings of group-based guilt. The low levels of all three negative group-based emotions signal that even though incivility might influence some emotions, White participants on average are not moved enough by the racist incident they encountered to feel strong group-based anger, guilt, or shame

While the civility manipulation did alter group-based anger and shame, seeing the exchange between two White people and seeing it between a White and a Black person, whether it was neutral or uncivil, had no effect on the White participants' negative emotions. Though the exchange in every condition was by definition racist because it involved a White person upholding the institutions that oppress Black people, the conditions featuring a Black victim were expected to augment the intergroup racism of the conversation, leading to more negative group-based emotions. It is possible that, because the conversation was about a racialized, political topic in every condition, that its racist nature was already salient enough to participants even in the exchange between two White people, and that the presumption about increased intergroup awareness is unfounded. It is also possible that, even if the intergroup salience did increase and the White participants acknowledged the racism more when the Black person was getting attacked, they just did not care enough to get upset. This relative apathy aligns with past research that finds people have very limited reactions of being upset when witnessing racism (Karmali et al., 2017).

Despite the race manipulation not influencing any negative emotions or action intentions, the uncivil manipulation was able to increase political action and comment response intentions via increased group-based anger. This quick, fiery emotion was the mobilizing catalyst needed for White

participants to increase willingness for online and offline racial justice actions. Group-based shame did not have a similar effect as anger on comment response intentions, likely because shame creates a sense of wanting to distance oneself from the shameful ingroup member instead of doing anything to repair the relationship with the victim (Lickel et al., 2005). However, the indirect effect of shame was found to be not significantly different than that of anger in predicting political action intentions, suggesting that shame might help catalyze political action intentions and warranting future investigation into the role of group-based shame in this context. Because of the other differential effects, though, it is clear that instead of grouping these emotions together as general negative affect, they should be treated as discrete in research pertaining to the defensive effect. While anger is no guarantee for bystander intervention (Dickter & Newton, 2013), other negative emotions might make users feel uncomfortable about hateful online comments without propelling action, thereby maintaining the racial status quo. Averaging them all together would dilute the effects of the emotions that are actually at work and potentially obfuscate important findings of online incivility research.

These mediation results support the notion that the defensive effect can be translated to intergroup contexts, in which the focus is on the social instead of the personal sense of self. We add another layer of complexity to typical defensive effect studies and highlight the utility of this phenomenon to circumstances that cannot simply be treated as interpersonal. This study finds that social identities and group-based emotions are akin to individual identities and personal emotions in creating the necessary conditions needed for online racist incivility to mobilize some White people for anti-racist action. Other White bystanders, however, while perhaps feeling some negative emotions at the remarks, intended to remain inactive and therefore perpetuate racism.

Limitations and future directions

Notwithstanding the contributions of this study, there are a few limitations to consider. First, along with race, participants shared other social identities (university affiliation, gender) with the commenters and could have also identified with commenters based on similar issue positions. Given that we did not assess which social identities were activated, it is possible that the observed effects are due to any one of these shared social identities. However, research does suggest that in a situation not directly involving the self, those experiencing emotions do so because of an activated social identity connection (Mackie et al., 2016). Further, while we cannot rule out the possibility that some participants felt anger due to a social identity connection with the person supporting the protests, group-based guilt and shame are evoked

from an ingroup member's wrongdoing, and the person disagreeing with the protests was the only one, especially in the uncivil conditions, with obvious wrongdoing. Despite this uncertainty about with whom the participants most identified, the results of this experiment still demonstrate the utility of the defensive effect in situations of heightened social identity salience since experiencing group-based anger mobilized participants to take action. Future work should directly measure social identity activation to determine which specific identity is useful to trigger group-based emotions and action intentions upon witnessing racist interactions.

Further, while the stimuli were constructed to look like realistic tweets, the process of viewing a static photo of a Twitter conversation in an online survey is different than seeing these kinds of comments naturally on Twitter, limiting ecological validity. It is possible that Twitter users might never stumble upon this kind of exchange, especially one between total strangers, and if they do, it might be among other tweets discussing the same issue. Future research should expose participants to a more natural Twitter setting to see if witnessing racist incivility in that environment produces similar results.

Also, while the comment response and political action intention measures provide useful information, it is unknown whether participants would actually engage in these racial justice actions. In the future, especially accompanying a more realistic Twitter environment, it would be advantageous to allow participants to click "like" or write out a reply to measure actual comment responses.

Because this study used an unrepresentative student-based convenience sample, conclusions from this experiment should not be over-generalized. Future research should replicate these results on a representative sample and see if any negative group-based emotions besides those tested here are relevant for this defensive effect. Anxiety, which has been shown to be mobilizing in a personally relevant incivility context (Lu & Myrick, 2016), could also potentially be translated into an intergroup context and used as a group-based emotion. Testing other discrete emotions would further provide evidence as to which emotions in particular mobilize action upon witnessing incivility.

Conclusion

This experiment makes important theoretical contributions in extending the defensive effect to an intergroup setting. Racist comments are likely to persist on social media, so it is helpful to know that some White people who witness uncivil racist comments will become angry and willing to take action for racial justice. However, more needs to be done to motivate White individuals to get angry about and stand up against pernicious racist comments that might not contain derogatory language but are still harmful. Increasing the amount of group-based anger that White bystanders feel could increase their willingness

to take racial justice action – action that could slowly lead to less racist online comments and a more equitable online and offline environment.

Data availability

The data that support the findings of this study are openly available in the Open Science Framework, at https://osf.io/rhv38/?view_only=86d6a2f8be8148e1bb9996fc7459757c.

Disclosure statement

The authors have received no financial or other benefit from this research. The authors declare that there is no conflict of interest regarding the publication of this article.

Notes on contributors

Jessica Roden (B.A., Vassar College) is a doctoral candidate in the Department of Communication and Media, University of Michigan. Her research focuses on perceptions of race and gender activism on social media.

Dr. Muniba Saleem is an associate professor in the Department of Communication at University of California, Santa Barbara and a faculty affiliate at the Institute of Social Research at University of Michigan. Dr. Saleem examines how media influences intergroup relations between racial, ethnic, and religious groups using social scientific methods.

ORCID

Jessica Roden http://orcid.org/0000-0003-1114-1354
Muniba Saleem http://orcid.org/0000-0002-6610-9880

References

Barnidge, M., Kim, B., Sherrill, L. A., Luknar, Ž., & Zhang, J. (2019). Perceived exposure to and avoidance of hate speech in various communication settings. *Telematics and Informatics*, *44*(101263), 1–13. https://doi.org/10.1016/j.tele.2019.101263

Borah, P. (2014). Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research*, *41*(6), 809–827. https://doi.org/10.1177/0093650212449353

Boycoff, J., & Carrington, B. (2020). Sporting dissent: Colin Kaepernick, NFL activism, and media framing contests. *International Review for the Sociology of Sport*, *55*(7), 829–849. https://doi.org/10.1177/1012690219861594

Brown, R., Gonźalez, R., Zagefka, H., Manzi, J., & Čehajić, S. (2008). Nuestra Culpa: Collective guilt and shame as predictors of reparation for historical wrongdoing. *Journal of Personality and Social Psychology*, *94*(1), 75–90. https://doi.org/10.1037/0022-3514.94.1.75

Casteel, K. (2017, October 9). How do Americans feel about the NFL protests? It depends on how you ask. *FiveThirtyEight*. https://fivethirtyeight.com/features/how-do-americans-feel-about-the-nfl-protests-it-depends-on-how-you-ask/

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Comenetz, J. (2016). *Demographic aspects of surnames from census 2010*. https://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf

DiAngelo, R. (2018). *White fragility: Why it's so hard for white people to talk about racism*. Beacon Press.

Dickter, C. L., & Newton, V. A. (2013). To confront or not to confront: Non-targets' evaluations of and responses to racist comments. *Journal of Applied Social Psychology*, *43*(S2), 262–275. https://doi.org/10.1111/jasp.12022

Doosje, B., Branscombe, N. R., Spears, R., & Manstead, A. S. R. (1998). Guilt by association: When one's group has a negative history. *Journal of Personality and Social Psychology*, *75*(4), 872–886. https://doi.org/10.1037/0022-3514.75.4.872

Duggan, M. (2017, July 11). Online harassment 2017. *Pew Research Center*. http://www.pewinternet.org/2017/07/11/online-harassment-2017/

Gaertner, S. L., & Dovidio, J. F. (1977). The subtlety of white racism, arousal, and helping behavior. *Journal of Personality and Social Psychology*, *35*(10), 691–707. https://doi.org/10.1037/0022-3514.35.10.691

Gordijn, E. H., Wigboldus, D., & Yzerbyt, V. (2001). Emotional consequences of categorizing victims of negative outgroup behavior as ingroup or outgroup. *Group Processes & Intergroup Relations*, *4*(4), 317–326. https://doi.org/10.1177/1368430201004004002

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

Hogg, M. A., & Reid, S. A. (2006). Social identity, self-categorization, and the communication of group norms. *Communication Theory*, *16*(1), 7–30.

Hwang, H., Kim, Y., & Kim, Y. (2018). Influence of discussion incivility on deliberation: An examination of the mediating role of moral indignation. *Communication Research*, *45*(2), 213–240. https://doi.org/10.1177/0093650215616861

Johns, M., Schmader, T., & Lickel, B. (2005). Ashamed to be an American? The role of identification in predicting vicarious shame for anti-Arab prejudice after 9-11. *Self and Identity*, *4*(4), 331–348. https://doi.org/10.1080/15298860500145822

Karmali, F., Kawakami, K., & Page-Gould, E. (2017). He said what? Physiological and cognitive responses to imagining and witnessing outgroup racism. *Journal of Experimental Psychology*, *146*(8), 1073–1085. https://doi.org/10.1037/xge0000304

Karmali, F., Kawakami, K., Vaccarino, E., Williams, A., Phills, C., & Friesen, J. P. (2019). I don't see race (or conflict): Strategic descriptions of ambiguous negative intergroup contexts. *Journal of Social Issues*, *75*(4), 1002–1034. https://doi.org/10.1111/josi.12353

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

Kim, J. W. (2018). Online incivility in comment boards: Partisanship matters–But what I think matters more. *Computers in Human Behavior*, *85*, 405–412. https://doi.org/10.1016/j.chb.2018.04.015

Kim, Y. M. (2009). Issue publics in the new information environment: Selectivity, domain specificity, and extremity. *Communication Research*, *36*(2), 254–284. https://doi.org/10.1177/0093650208330253

Kuppens, P., Van Mechelen, I., Smits, D. J. M., De Boeck, P., & Ceulemans, E. (2007). Individual differences in patterns of appraisal and anger experience. *Cognition and Emotion*, *21*(4), 689–713. https://doi.org/10.1080/02699930600859219

Lane, D. S., Coles, S. M., & Saleem, M. (2019). Solidarity effects in social movement messaging: How cueing dominant group identity can increase movement support. *Human Communication Research*, *45*(1), 1–26. https://doi.org/10.1093/hcr/hqy011

Leach, C. W., Iyer, A., & Pederson, A. (2006). Anger and guilt about ingroup advantage explain the willingness for political action. *Personality and Social Psychology Bulletin*, *32*(9), 1232–1245. https://doi.org/10.1177/0146167206289729

Lee-Won, R. J., Lee, J. Y., Song, S., & Borghetti, L. (2017). "To the bottle I go … to drain my strain": Effects of microblogged racist messages on target group members' intention to drink alcohol. *Communication Research*, *44*(3), 388–415. https://doi.org/10.1177/0093650215607595

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*, *81*(1), 146–159. https://doi.org/10.1037/0022-3514.81.1.146

Lickel, B., Schmader, T., Curtis, M., Scarnier, M., & Ames, D. R. (2005). Vicarious shame and guilt. *Group Processes & Intergroup Relations*, *8*(2), 145–157. https://doi.org/10.1177/1368430205051064

Lu, Y., & Myrick, J. G. (2016). Cross-cutting exposure on Facebook and political participation: Unraveling the effects of emotional responses and online incivility. *Journal of Media Psychology*, *28*(3), 100–110. https://doi.org/10.1027/1864-1105/a000203

Mackie, D. M., Maitner, A. T., & Smith, E. R. (2016). Intergroup emotions theory. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 149–174). Psychology Press.

Mackie, D. M., Devos, T., & Smith, E. R. (2000). Intergroup emotions: Explaining offensive action tendencies in an intergroup context. *Journal of Personality and Social Psychology*, *79*(4), 602–616. https://doi.org/10.1037/0022-3514.79.4.602

Mallett, R. K., Huntsinger, J. R., Sinclair, S., & Swim, J. K. (2008). Seeing through their eyes: When majority group members take collective action on behalf of an outgroup. *Group Processes & Intergroup Relations*, *11*(4), 451–470. https://doi.org/10.1177/1368430208095400

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*(1), 37–68. https://doi.org/10.1080/14792779543000011

Marques, J. M., Yzerbyt, V. Y., & Leyens, J. P. (1988). The "black sheep effect": Extremity of judgments towards ingroup members as a function of group identification. *European Journal of Social Psychology*, *18*(1), 1–16. https://doi.org/10.1002/ejsp.2420180102

Masullo Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Palgrave Macmillan.

Masullo Chen, G. M., Fadnis, D., & Whipple, K. (2020). Can we talk about race? Exploring online comments about race-related shootings. *Howard Journal of Communications*, *31*(1), 35–49. https://doi.org/10.1080/10646175.2019.1590256

Masullo Chen, G. M., & Lu, S. (2017). Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website

comments. *Journal of Broadcasting & Electronic Media*, *61*(1), 108–125. https://doi.org/10.1080/08838151.2016.1273922

Montepare, J. M., & Opeyo, A. (2002). The relative salience of physiognomic cues in differentiating faces: A methodological tool. *Journal of Nonverbal Behavior*, *26*(1), 43–59. https://doi.org/10.1023/A:1014470520593

Mutz, D. C. (2015). *In-your-face politics: The consequences of uncivil media*. Princeton University Press.

Nabi, R. L. (2010). The case for emphasizing discrete emotions in communication research. *Communication Monographs*, *77*(2), 153–159. https://doi.org/10.1080/03637751003790444

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

Norris, P. (2007). Political activism: New challenges, new opportunities. In C. Boix & S. Stokes (Eds.), *The Oxford handbook of comparative politics* (pp. 628–652). Oxford University Press.

Norton, M. I., Sommers, S. R., Apfelbaum, E. P., Pura, N., & Ariely, D. (2006). Color blindness and interracial interaction: Playing the political correctness game. *Psychological Science*, *17*(11), 949–953. https://doi.org/10.1111/j.1467-9280.2006.01810.x

Otto, L. P., Lecheler, S., & Schuck, A. R. T. (2020). Is context the key? The (non-) differential effects of mediated incivility in three European countries. *Political Communication*, *37*(1), 88–107. https://doi.org/10.1080/10584609.2019.1663324

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Puryear, C., & Vandello, J. A. (2019). Inflammatory comments elicit less outrage when made in anonymous online contexts. *Social Psychological and Personality Science*, *10*(7), 895–902. https://doi.org/10.1177/1948550618806350

Rains, S. A., Kenski, K., Coe, K., & Harwood, J. (2017). Incivility and political identity on the internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication*, *22*(4), 163–178. https://doi.org/10.1111/jcc4.12191

Saleem, M., Hawkins, I., Wojcieszak, M. E., & Roden, J. (2021). When and how negative news coverage empowers collective action in minorities. *Communication Research*, *48*(2), 291–316. https://doi.org/10.1177/0093650219877094

Saucier, D. A., Miller, C. T., & Doucet, N. (2005). Differences in helping whites and blacks: A meta-analysis. *Personality and Social Psychology Review*, *9*(1), 2–16. https://doi.org/10.1207/s15327957pspr0901_1

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Scott, J. P. (1980). The function of emotions in behavioral systems: A systems theory analysis. In R. Plutchik & H. Kellerman (Eds.), *Emotions: Theory, research, experience* (pp. 35–56). Academic Press.

Seger, C. R., Smith, E. R., & Mackie, D. M. (2009). Subtle activation of a social categorization triggers group-level emotions. *Journal of Experimental Social Psychology*, *45*(3), 460–467. https://doi.org/10.1016/j.jesp.2008.12.004

Shepherd, L., Spears, R., & Manstead, A. S. R. (2013). 'This will bring shame on our nation': The role of anticipated group-based emotions on collective action.

Journal of Experimental Social Psychology, *49*(1), 42–57. https://doi.org/10.1016/j.jesp.2012.07.011

Smith, E. R., Seger, C. R., & Mackie, D. M. (2007). Can emotions be truly group level? Evidence regarding four conceptual criteria. *Journal of Personality and Social Psychology*, *93*(3), 431–446. https://doi.org/10.1037/0022-3514.93.3.431

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Spears, R. (2011). Group identities: The social identity perspective. In S. J. Schwartz, K. Luyckx, & V. L. Vignoles (Eds.), *Handbook of identity theory and research* (pp. 201–224). Springer.

Swim, J. K., & Miller, D. L. (1999). White guilt: Its antecedents and consequences for attitudes toward affirmative action. *Personality and Social Psychology Bulletin*, *25*(4), 500–514. https://doi.org/10.1177/0146167299025004008

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks-Cole.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, L. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1–38. https://doi.org/10.18637/jss.v059.i05

Van Zomeren, M., Postmes, T., & Spears, R. (2008). Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin*, *134*(4), 504–535. https://doi.org/10.1037/0033-2909.134.4.504

Van Zomeren, M., Postmes, T., Spears, R., & Bettache, K. (2011). Can moral convictions motivate the advantaged to challenge social inequality? Extending the social identity model of collective action. *Group Processes & Intergroup Relations*, *14*(5), 735–753. https://doi.org/10.1177/1368430210395637

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research*, *47*(6), 921–944. https://doi.org/10.1177/0093650219855330