Check for updates

Article

Communication Research
1–32
© The Author(s) 2026

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00936502261489660
journals.sagepub.com/home/crx


From Hate Speech to Collective Action Intentions: The Role of Counter-Speech and Accuracy Cues Among Targeted Minorities



Desiree Schmuck[1], Magdalena Obermaier[2], Nancy Molina-Rogers[3], and Muniba Saleem[4]

Abstract

Online hate speech targeting minority groups poses a significant social identity threat to those affected, often prompting coping strategies like out-group avoidance or derogation, which can harm intergroup relations in diverse societies. However, collective action as an empowering response has received limited attention in this context. Drawing on the social identity model of collective action (SIMCA), this study examines both the mechanisms and boundary conditions of online hate speech's impact on minority groups. Across two experiments (N_{Study1} = 328 Muslims; N_{Study2} = 583 Latinos), we find that online hate speech increases group-based injustice, which in turn drives collective action intentions, supporting the SIMCA. This effect is amplified when information embedded in the online hate speech is perceived as inaccurate or flagged for inaccuracy. In Study 2, we further demonstrate that counter-speech can reduce group-based injustice and, when paired with low-accuracy hate speech, enhance collective efficacy beliefs.

[1]University of Vienna, Austria
[2]LMU Munich, Germany
[3]University of California, Los Angeles, USA
[4]University of California, Santa Barbara, USA

Corresponding Author:
Desiree Schmuck, Department of Communication, University of Vienna, Währinger Straße 29, Vienna 1090, Austria.
Email: desiree.schmuck@univie.ac.at

Keywords
online hate speech, counter-speech, social identity, collective action, minorities

Introduction

Harmful communication is a pervasive feature of online environments. A growing body of research has examined such interactions under the broader concept of online incivility, which encompasses communicative behaviors ranging from subtle violations of respectful discourse to overtly rude or dismissive language (Pradel et al., 2024; Ziegele et al., 2020). Empirical evidence suggests that uncivil communication constitutes a substantial share of online discourse, particularly in political contexts (Theocharis et al., 2020). Exposure to incivility has been linked to negative evaluations of social groups (Kim, 2024), increased attitude polarization (Anderson et al., 2014), and reduced political trust, although its effects on participation remain mixed (Van't Riet & Van Stekelenburg, 2022).

Among the more severe forms of harmful online communication is online hate speech, which refers to communication that attacks, derogates, or intimidates individuals based on group-related characteristics such as gender, ethnicity, or national origin (Hawdon et al., 2017; Warner & Hirschberg, 2012). Social media platforms, due to their public visibility and perceived anonymity, facilitate its widespread dissemination (Dredge et al., 2014). Cross-national research indicates that 43% of users have encountered hate speech online (Hawdon et al., 2017). Existing research has examined the effects of harmful online speech among both bystanders (e.g., Pabian, 2025) and members of targeted groups (e.g., Van Houtven et al., 2024). Yet, the literature has disproportionately focused on bystanders, who constitute the largest share of online users. While prior research on online incivility has largely focused on general audiences or bystanders, these approaches do not capture the identity-relevant processes triggered when individuals are directly targeted. Consequently, despite growing knowledge about the societal consequences of uncivil online discourse, we still know relatively little about how individuals directly targeted by group-based hate interpret, psychologically process, and behaviorally respond to these experiences.

Among targets, exposure to hate speech has been associated with severe psychological consequences, including effects comparable to other traumatic experiences (Leets, 2002) and impairments in well-being such as heightened anxiety or sleep disturbances (Delgado & Stefancic, 2019). Beyond these individual-level harms, being directly targeted by group-based online hate differs fundamentally from merely witnessing uncivil discourse. Direct targeting is more likely to activate social identity concerns, heighten perceptions of group-based threat and injustice, and elicit identity-relevant coping responses among minority group members (Branscombe et al., 1999; Major & O'Brien, 2005; van Zomeren, Postmes, et al., 2008). As a result, the psychological and behavioral processes triggered by direct victimization cannot be readily inferred from bystander-oriented research.

Specifically, we identify the following research gaps: First, while existing studies have primarily focused on psychological outcomes like loneliness (e.g., Stahel & Baier, 2023) or stress (Shmulewitz et al., 2025) as well as intergroup attitudes and relations as outcomes of hate speech toward minorities (Obermaier et al., 2023; Schäfer et al., 2024; Van Houtven et al., 2024), little attention has been devoted to the mobilizing potential of online hate speech. According to the social identity model of collective action (SIMCA, van Zomeren, Postmes, et al., 2008), an activated social identity increases collective action, defined as group-based efforts aimed at improving a group's social standing (van Zomeren, Postmes, et al., 2008), via group-based injustice and/or collective efficacy. Collective action may take normative forms, including joining protests or signing petitions, as well as non-normative forms, such as aggressive or violent behaviors that may be detrimental to society (van Zomeren, Postmes, et al., 2008). Investigating collective action as a coping response to online hate speech is crucial, given the pervasive visibility of such content and difficulty of avoiding it in digital environments (Hawdon et al., 2017).

Second, key boundary conditions remain underexamined. Counter-speech, which is a communicative practice through which individuals verbally oppose hate speech (Buerger & Wright, 2019; McDonald et al., 2016), can mitigate perceived identity threat (Van Houtven et al., 2024) and reduce hostile reactions (Obermaier et al., 2023). Yet, its effects may depend on the counter-speaker's group membership. Similarly, accuracy cues may shape variations in perceptions of legitimacy or social credibility of derogatory claims, thereby influencing group-based injustice (Saleem et al., 2021).

To address these gaps, we conducted two online experiments with minority group members directly targeted by hate speech. Study 1 focuses on Muslims in Germany, a frequent target of online hostility (LfM, 2024), while Study 2 examines racist hate speech toward Latinos in the United States, a population that has increasingly been the target of xenophobia (The State of Hate Report, 2024). Across studies, we test whether exposure to online hate speech increases collective action intentions via group-based injustice and collective efficacy. We further test whether these effects are moderated by counter-speech and accuracy cues. This research provides important insight into when and how targeted individuals respond to online hate speech by distinguishing between normative and non-normative collective action, which have distinct implications for legitimacy, public support, and the potential escalation of social change (Becker & Tausch, 2015).

The Social Identity Model of Collective Action (SIMCA)

The Social Identity Model of Collective Action (van Zomeren, Postmes, et al., 2008) suggests that social identities provide the foundation for a process in which individuals belonging to disadvantaged groups develop a sense of agency, ultimately leading to collective action. SIMCA specifically posits that individuals are driven to engage in collective action due to perceptions of group-based injustice or collective efficacy (van Zomeren, Postmes, et al., 2008). *Injustice* refers to the "subjective experience of unjust disadvantage," which arises from appraisal processes that integrate

both evaluative and affective components (van Zomeren, Postmes, et al., 2008). In intergroup contexts, van Zomeren et al. (2018, p. 123) suggest using the umbrella term of "group-based injustice," which, they argue, may refer, in specific studies, to measures of group-based relative deprivation, perceptions of group-based injustice, or the experience of group-based anger. Experiences of group-based injustice motivate collective action intentions, as individuals seek to challenge what they experience as unfair treatment of their group. Collective efficacy, on the other hand, is the shared belief in the group's ability to change its circumstances and future. When individuals hold stronger beliefs in collective efficacy, they are more motivated to participate in collective action, as they believe their efforts will lead to achieving relevant goals (van Zomeren, Postmes, et al., 2008).

Online Hate Speech and Collective Action

Existing research has approached the effects of harmful online speech in different ways, depending on whether individuals are exposed as bystanders or are directly targeted. Beyond research explicitly focusing on hate speech, a growing body of scholarship has examined the effects of harmful online communication more broadly including research on online incivility (e.g., Ziegele et al., 2020). In prior research, incivility has been conceptualized as covering a range of communicative behaviors, from subtle violations of respectful discourse to clearly rude or dismissive language (Pradel et al., 2024). Empirical studies show that uncivil communication constitutes a substantial share of political discourse online. For instance, analyses of Twitter interactions with legislators indicate that uncivil messages account for roughly one-fifth of all related tweets, with noticeable surges during contentious policy debates and major political events. While some instances reflect coordinated campaigns, uncivil language is largely produced by a broad and diverse set of users rather than a small group of actors (Theocharis et al., 2020).

While much of this research examines incivility as a general communicative phenomenon, an important distinction concerns whether it explicitly references social groups or identities. References to social groups within media environments, including digital media, serve as important situational cues that heighten the salience of social identity (Major & O'Brien, 2005). If negative or stereotypical, these depictions can undermine motivations for positive group distinctiveness and threaten social identity (Saleem & Ramasubramanian, 2019). Such identity threats are especially likely in the context of online hate speech (Van Houtven et al., 2024), which inherently targets individuals based on group membership (Hawdon et al., 2017; Warner & Hirschberg, 2012). Compared to more indirect or ambiguous negative portrayals of social groups in news media (Saleem et al., 2019), hate speech involves explicit hostility, derogation, and intentional targeting of social groups. It often occurs in interactive environments where individuals are directly addressed (Bilewicz & Soral, 2020), thereby intensifying its psychological impact (Shmulewitz et al., 2025). Hate speech frequently relies on dehumanizing metaphors—such as referring to groups as "parasites" or "animals," and may include direct threats or incitements to violence (Paasch-Colberg

et al., 2021). As a result, exposure to hate speech constitutes a more severe and unambiguous violation of group-based norms compared to negative news portrayals, posing a direct threat to social identity and heightening perceptions of group-based injustice among targeted group members (Obermaier et al., 2023; Van Houtven et al., 2024).

Applying SIMCA to this context suggests that online hate speech may contain many aspects that can motivate collective action intentions. Exclusionary narratives delineate boundaries between favored ingroups and disparaged outgroups, increasing awareness of group-based disadvantage and heightening group-based injustice. At the same time, shared threats can strengthen group unity and beliefs in the group's collective capacity to respond, or collective efficacy (Drury & Reicher, 2009). Together, these processes suggest that exposure to online hate speech may activate the key mechanisms identified in SIMCA (group-based injustice and collective efficacy) that motivate collective action intentions.

Despite the theoretical connections, existing research on online hate speech has not directly tested the roles of group-based injustice and collective efficacy in motivating collective action strategies. However, related studies have found evidence for closely linked mechanisms. For instance, prior research suggests that online hate speech can provoke emotions tied to group-based injustice, such as anger (Boeckmann & Liew, 2002; Schmid et al., 2024). Additionally, research on the impact of negative portrayals of one's ingroup in the media has shown that such exposure can enhance collective efficacy beliefs, which in turn motivate collective action strategies (Saleem et al., 2021). Drawing on SIMCA and this prior evidence, we hypothesize that:

H1: Online hate speech in comparison to no hate speech increases (a) group-based injustice and (b) collective efficacy, which subsequently enhance collective action intentions.

Counter-Speech

Although exposure to online hate speech may mobilize collective responses, these effects are unlikely to occur uniformly. Contextual cues embedded within online interactions, such as the presence of counter-speech, may shape how targets interpret identity threats and whether mobilizing psychological processes are strengthened or attenuated. Counter-speech refers to communicative efforts by individuals or groups—whether targets or bystanders—to publicly challenge hateful discourse to reduce its impact (Buerger & Wright, 2019; McDonald et al., 2016). By signaling social norms that oppose hate, counter-speech may alter perceptions of injustice and collective efficacy that drive collective action responses. While counter-speech may facilitate collective processes, it does not constitute collective action per se, as it primarily refers to *communicative* efforts aimed at challenging, correcting, or resisting harmful content, often taking the form of direct replies or public expressions of disagreement (Buerger & Wright, 2019). In contrast, collective action involves broader coordinated efforts by groups of individuals to achieve broader social or political change, such as organizing

campaigns, petitions, or protests (van Zomeren, Postmes, et al., 2008). However, in some cases, counter-speech can become a formally organized activity, for example in the form of a social movement, which is aimed at collectively countering uncivil comments (Ziegele et al., 2020). From a social identity perspective, these dynamics are particularly relevant because responses to hate speech are shaped by perceived identity threats, which can translate into group-based injustice and motivate collective action.

Studies suggest that counter-speech can diminish the perceived identity threat posed by hate speech (Van Houtven et al., 2024). Although not yet directly examined, counter-speech may mitigate group-based injustice by challenging threats and signaling social norms that reaffirm group status, thus reducing identity threats and restoring fairness (Tajfel & Turner, 1986).

From a social identity perspective, the implications of counter-speech for group-based injustice are theoretically ambiguous. On the one hand, counter-speech may challenge hostile narratives and signal that hateful messages are socially unacceptable, thereby reducing identity threat and dampening grievance. On the other hand, by explicitly acknowledging the violation and framing it as illegitimate, counter-speech may validate targets' perceptions that an injustice has occurred, reinforcing rather than attenuating perceived grievance (Bilewicz & Soral, 2020; van Zomeren, Postmes, et al., 2008). These effects may further depend on the identity of the counter-speaker and whether they are part of one's ingroup or not (Obermaier et al., 2023; Van Houtven et al., 2024). These contrasting mechanisms suggest that the effects of counter-speech on group-based injustice are theoretically unclear leading to the following research question:

RQ1: How does counter-speech by (a) a minority group member and (b) a majority group member influence group-based injustice and subsequent collective action intentions in response to online hate speech?

In contrast to its ambiguous effects on group-based injustice, counter-speech is more consistently expected to strengthen collective efficacy. Observing others publicly challenge hate speech may signal that resistance is possible and socially supported, fostering beliefs that collective efforts can successfully address group-based disadvantage (Outten & Schmitt, 2015; van Zomeren, Postmes, et al., 2008). By demonstrating that group members or allies are willing to intervene, counter-speech may enhance perceptions of collective agency and thereby increase collective action intentions.

H2: Counter-speech increases collective efficacy, which in turn enhances collective action intentions.

Beyond these main effects, the effectiveness of counter-speech may depend on the group membership of the counter-speaker. Counter-speech can be expressed by individuals outside the targeted group, often representing the majority, referred to as

majority counter-speech. Conversely, *minority counter-speech* is delivered by someone who is also a target of online hate speech, that is, another minority group member (Obermaier et al., 2023; Van Houtven et al., 2024). Prior research suggests that the group membership influences the extent to which counter-speech can mitigate the negative effects of online hate speech. For example, Van Houtven et al. (2024) found that majority counter-speech reduced the social identity threat posed by online hate speech. These effects have been attributed to reductions in aggressive cognitions, emotions, and behavioral tendencies to respond defensively on behalf of one's ingroup (Hsueh et al., 2015). In this context, counter-speech delivered by majority-group members may carry stronger normative signaling power because it indicates that dominant-group actors reject the hostile message, whereas minority-group counter-speech may be perceived as ingroup defense. However, Obermaier et al. (2023) showed that counter-speech by both majority and minority members reduced hateful responses. Hence, prior evidence comparing the effects of majority and minority counter-speech is mixed (Obermaier et al., 2023; Van Houtven et al., 2024), suggesting the need for further investigation. Therefore, we ask:

RQ2: Does the group membership of the counter-speaker (majority vs. minority) moderate the effects of counter-speech on (a) group-based injustice and (b) collective efficacy?

The Moderating Role of (Perceived) Accuracy

Beyond counter-speech, the perceived credibility or accuracy of the claims embedded in hate speech may shape how targeted individuals interpret identity threats and respond to them. We argue that such accuracy cues primarily operate through perceptions of legitimacy. Specifically, accuracy cues signal whether derogatory statements are perceived as socially credible or legitimate, thereby influencing the key psychological mechanisms identified in the Social Identity Model of Collective Action (SIMCA; van Zomeren, Postmes, et al., 2008). Although prior research has examined the role of accuracy in domains such as misinformation and uncivil online discourse (Masullo et al., 2023; Pennycook et al., 2018), its influence on reactions to online hate speech, particularly among targeted group members, remains largely unexplored.

Hate speech often combines derogatory rhetoric with misleading or false claims about social groups (Hameleers et al., 2022). When such claims are perceived or flagged as inaccurate, they may reinforce the perception that group-based derogation is illegitimate, heightening group-based injustice and strengthening motivations to challenge the message collectively. Conversely, when hate speech is perceived as accurate, negative portrayals may be interpreted as reflecting legitimate social realities, potentially reducing group-based injustice and dampening collective action motivations. According to the System Justification Theory (Jost & Banaji, 1994), marginalized individuals may, under certain conditions, come to accept negative portrayals as reflecting legitimate social arrangements, thereby weakening both grievance and the perceived necessity of collective resistance (Jost et al., 2017).

Although hate speech is likely to be perceived as inaccurate by most targets, digital environments may reduce sensitivity to derogatory claims and alter perceptions of its normative acceptance (Bilewicz & Soral, 2020). Consequently, even subtle variations in perceived accuracy of claims made in hateful comments may signal legitimacy and shape responses (Bettencourt et al., 2001). Importantly, perceived accuracy in this context does not merely reflect agreement with hateful claims but captures the extent to which such claims are viewed as socially credible or reflective of broader societal beliefs. Even when individuals reject hate speech as false, variation in perceived credibility may signal whether derogatory narratives are interpreted as fringe rhetoric or as representing widely shared societal attitudes (Bettencourt et al., 2001). These legitimacy appraisals are consequential because they shape whether group-based disadvantage is perceived as unjust and contestable or as socially entrenched and difficult to change. Accordingly, both subjective perceptions of accuracy and platform-provided accuracy cues are expected to moderate the extent to which exposure to hate speech influences group-based injustice, collective efficacy, and collective action intentions. Supporting this reasoning, prior research shows that when negative portrayals of minority groups are perceived as inaccurate, individuals report stronger group-based injustice, greater collective efficacy, and increased willingness to engage in collective action (Saleem et al., 2021). Thus, we hypothesize:

H3: Greater perceived or platform-verified accuracy of the information embedded in online hate speech decreases (a) group-based injustice and (b) collective efficacy, which are subsequently related to collective action intentions.

Study 1

Method

To test our hypotheses and RQs (see Figure 1), we carried out an online experiment comprising four experimental conditions: (a) the presence of standalone Islamophobic online hate speech without counter-speech, (b) online hate speech countered by a member of the majority group (a non-Muslim), (c) online hate speech countered by another minority group member (a Muslim), and (d) a control group without online hate speech or counter-speech (see also Obermaier et al., 2023).

We collected the data in May 2020 with the help of a survey company. The company recruited individuals residing in Germany who identified Turkish and/or Arabic as their native languages, resulting in a total of 426 participants. The sample was adjusted to exclude participants who did not identify Islam as their religion, leading to a sample of $N = 362$ Muslims. Additionally, participants with a retention time of one standard deviation below the mean dwell time ($\leqslant 304$ s) and those reporting significant difficulty understanding the questionnaire were excluded, resulting in a final sample of $N = 328$ Muslims. All participants provided consent to participate and were provided with a trigger warning at the outset of the survey. Participants were informed

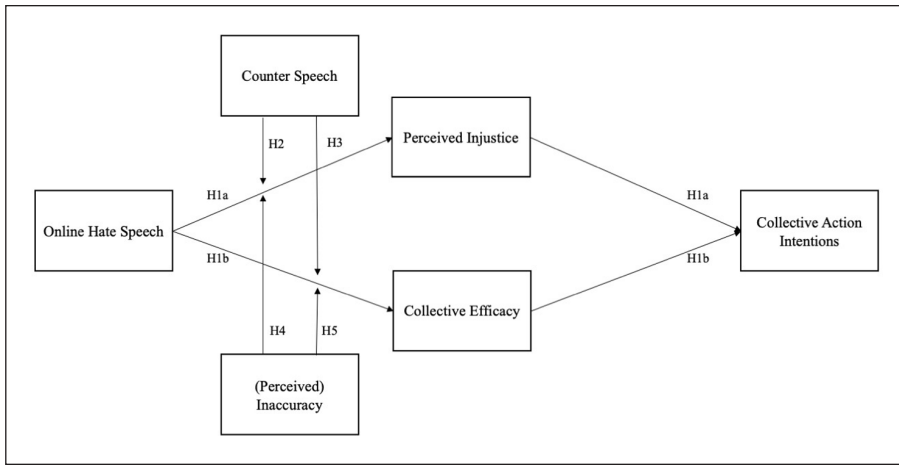


Figure 1. Hypothesized model.

of their right to terminate participation at any point if they felt uncomfortable. The survey concluded with a comprehensive debriefing, providing detailed information about the fictitious posts and comments, and a clear statement emphasizing that the hateful statements did not reflect the researchers' attitudes. Participation was voluntary, and participants received incentives. The study materials, dataset, and analysis script are available on OSF: https://osf.io/kghjy/overview.

Sample. Among the participants, 48% were female and 52% male, with an average age of 33 years ($SD = 9.80$). Furthermore, 68% possessed a high school degree. Regarding migration background, 34% had no migration background, 41% were born in Germany with at least one parent born in another country, 25% were born in another country themselves. Among those with migration background, 64% had resided in Germany for over 10 years, and 67% reported German as their native language. Finally, religious identification was $M = 3.70$ ($SD = 0.89$) on a five-point scale (1 = "*not religious*" to 5 = "*very religious*").

Randomization Check. A randomization check showed no significant variations among the experimental groups concerning gender, $\chi^2(3, 328) = 3.66$, $p = .30$; age, $F(3, 324) = 0.60$, $p = .61$, $\eta^2 = .01$; migration background, $\chi^2(6, 328) = 5.73$, $p = .45$; duration of residence in Germany, $F(3, 79) = 0.54$, $p = .66$, $\eta_{part} = .02$; prior national identification, $F(3, 324) = 1.28$, $p = .28$, $\eta^2 = .01$; and prior religious identification, $F(3, 324) = 0.54$, $p = .66$, $\eta^2 = .01$.

Stimulus Materials. Participants were exposed to a fictitious Facebook post shared by an online branch of a respected German newspaper, promoting an article titled "Several Daycare Centers eliminate Pork for Children" (see Supplemental Appendix B).

The post and the comment(s) were created in the Facebook design incorporating various features such as mouse-over effects to enhance realism.

The online hate speech and counter-speech content drew upon prior research (Leonhard et al., 2018; Naab et al., 2018), with comments maintained at comparable lengths of 90 to 100 words. A fictitious user "Alex" was the author of the first comment. In the control condition, this comment was neutral stating "Interesting! Muslim kids in day-care." The baseline comment was intended to be neutral in evaluative content, meaning that it did not include derogatory language, group-based threat, or explicit negative judgment. In contrast, in the treatment conditions, the comment featured Islamophobic hate speech including defamations ("[they] live in their small Muslim parallel world from our taxes") and dehumanization ("these parasites"). Moreover, the hate speech condemned the imposition of the values of the minority group onto the majority ("we have to pay for them and then let them tell us how to live") and made a diffuse threat of violence ("we must fight back"; Leader Maynard et al., 2016).

This comment was followed by counter-speech in the two counter-speech conditions. The counter-speech was represented by a second comment either by a majority member ("Mark") or another minority member ("Malik"). We clarified the religious affiliation by stating in the comment "I am/I am not a Muslim." The comment countered the hate speech with arguments, stating ("These are just insinuations against Muslims that are totally inappropriate and wrong."), while also flagging it as unacceptable. While counter-speech can take many forms, including providing facts or attacking perpetrators (Obermaier et al., 2023), the counter-speech used in this study can be classified as factual counter-speech that adheres to deliberative norms, contradicts hateful claims and demonstrates support for those attacked in the hate speech (Ziegele et al., 2020). Additionally, it invoked shared values and unity, referencing concepts like "open society" and "standing together" (Dovidio et al., 2007).

Measures. If not stated otherwise, all constructs were measured on five-point Likert scales (1 = *strongly disagree*, 5 = *strongly agree;* Supplemental Appendix A presents an overview of all items, while Table A3 presents the zero-order correlations). In Study 1, *group-based injustice* was operationalized as the affective experience of anger, which is a common way to operationalize the umbrella term group-based injustice (see e.g., van Zomeren, 2020; van Zomeren et al., 2018). We used two items asking participants how they felt while viewing the Facebook post introduced with the prompt "Please indicate how you felt while watching the content?" followed by the items "I felt angry," and "I felt furious" (Pearson $r = .77$, $M = 3.21$, $SD = 1.40$). We measured *collective efficacy* based on Saleem et al. (2021) and Glasford and Calcagno (2012) with three items introduced with "The next questions relate to actions that can contribute to a positive image of Muslims in Germany. Please indicate how much you agree with the following statements" followed by (1) "I think Muslims should approach non-Muslim Germans about their values and ideals," (2) "I think Muslims can do something together as a group to improve their situation in Germany," (3) "Muslims should work together more to achieve greater political influence in Germany"

(Cronbach's $\alpha = .70$, $M = 3.56$, $SD = 0.91$). Similar to Saleem et al. (2021), we gauged *perceived accuracy* with six items, that is, "(1) balanced," (2) "credible," (3) "trustworthy," (4) "competent," (5) "correct," (6) "true" (Cronbach's $\alpha = .96$, $M = 2.18$, $SD = 1.25$). *Collective action intentions* were measured with three items introduced with "Please also indicate how strongly you agree with the following statements" followed by (1) "I think Muslims should defend themselves in the social media against hateful non-Muslim Germans," (2) "I am motivated to join a group on social media that helps Muslims against hateful expressions," (3) "I am motivated to join a group on social media that opposes the expression and spread of hatred against minorities (e.g., #ichbinhier)" (Cronbach's $\alpha = .79$, $M = 3.31$, $SD = 1.05$). As the SIMCA also predicts that "identification with the disadvantaged group is the proximal predictor of collective action" (van Zomeren, Postmes, et al., 2008, p. 507), we also measured identification with the disadvantaged group as a manipulation check to evaluate whether individuals in the hate speech groups were more likely to see the group they identify with as degraded or disadvantaged using four items (i.e., "As Muslim I feel. . ." "discriminated," "degraded," "offended," and "threatened" ". . .by this content"; $\alpha = .92$, $M = 3.35$, $SD = 1.33$; Leonhard et al., 2018).

Manipulation Check

Our manipulation check confirmed the effectiveness of the manipulations. Participants perceived the group they identify with (i.e., Muslims) as more disadvantaged ($M = 3.58$) than the control group ($M = 2.63$), $t(124.23) = -5.51$, $p < .001$. Moreover, most participants correctly recalled whether they had been exposed to standalone hate speech (71%) or to hate speech accompanied by a counter comment (81%), $\chi^2(1) = 89.74$, $p < .001$. Participants in both counter-speech conditions indicated that the counter comment clearly contradicted the hate speech, with no differences between counter-speech attributed to a non-Muslim ($M = 4.09$, $SD = 1.16$) and a Muslim source ($M = 4.01$, $SD = 1.27$), $t(138) = 0.35$, $p = .73$. However, counter-speech attributed to a Muslim source was perceived as slightly more defending of Muslims ($M = 4.30$, $SD = 1.00$) than counter-speech attributed to a non-Muslim source ($M = 3.91$, $SD = 1.15$), $t(138) = -2.12$, $p = .036$. Finally, most participants correctly identified whether the counter-speech was posted by a non-Muslim (76%) or a Muslim (94%). Overall, these results indicate that the experimental manipulations functioned as intended (see Supplemental Appendix A for full list of items).

Data Analysis

We ran path analyses in the *lavaan* package (Rosseel, 2012). The experimental group variable was dummy coded. We first estimated a path model with the control group as reference category to test H1 and the corresponding indirect effects. To test H3, we estimated a separate moderation model in which perceived accuracy was mean-centered and interaction terms between perceived accuracy and each experimental-condition dummy variable were entered simultaneously. We tested RQ1 and H2 with

Table 1. Path Model Study 1 (Control Group is the Reference Group).

	Group-based injustice		Collective efficacy		Collective action	
Predictor	*b*	*SE*	*b*	*SE*	*b*	*SE*
Standalone HS	.88***	0.22	−.08	0.15	−.20	0.13
Majority CS	1.00***	0.22	−.03	0.14	−.33	0.14
Minority CS	1.11***	0.21	−.07	0.14	.01	0.13
Group-based Injustice					.10**	0.04
Collective Efficacy					.63***	0.06
R^2	.10		.001		.33	

Note. HS = Hate Speech, CS = Counter-speech.
$N = 328$, $^{**}p < .01$. $^{***}p < .001$.

standalone hate speech as the reference group and RQ2 with majority counter-speech as the reference group. We used 5,000 bootstrap samples to calculate percentile confidence intervals for the inference of indirect effects.

Results

The first path model (see Table 1 and Figure 2) showed that all treatment groups expressed higher group-based injustice than the control group (standalone hate speech: $b = .88$, $SE = 0.22$, $p < .001$; Muslim counter-speech: $b = 1.11$, $SE = 0.21$, $p < .001$; non-Muslim counter-speech: $b = 1.00$, $SE = 0.22$, $p < .001$), supporting H1a. Contrary to H1b, however, we found no statistically significant effects on collective efficacy (standalone hate speech: $b = -.08$, $SE = 0.15$, $p = .594$; Muslim counter-speech: $b = -.07$, $SE = 0.14$, $p = .596$; non-Muslim counter-speech: $b = -.03$, $SE = 0.14$, $p = .811$). Both group-based injustice ($b = .10$, $SE = 0.04$, $p = .009$) and collective efficacy ($b = .63$, $SE = 0.06$, $p < .001$) significantly predicted collective action intentions.

Our findings further showed significant indirect effects from all conditions on collective action via group-based injustice, but not via collective efficacy (see Table A1). Hence, H1a was supported, while H1b was rejected.

To test RQ1 and H2, we used the other set of dummy variables with the standalone hate speech as reference group in our second path model (see Table 2). We found no significant differences between standalone hate speech and hate speech accompanied by majority counter-speech on group-based injustice ($b = .12$, $SE = 0.21$, $p = .568$) or efficacy ($b = .04$, $SE = 0.15$, $p = .759$). We also found no difference between standalone hate speech and hate speech accompanied by minority counter-speech on group-based injustice ($b = .23$, $SE = 0.21$, $p = .282$) or efficacy ($b = .01$, $SE = 0.14$, $p = .961$). Thus, counter-speech by either majority or minority group members did not systematically alter group-based injustice (RQ1) or collective efficacy (H2). Consistent with these findings, none of the indirect effects of majority or minority counter-speech

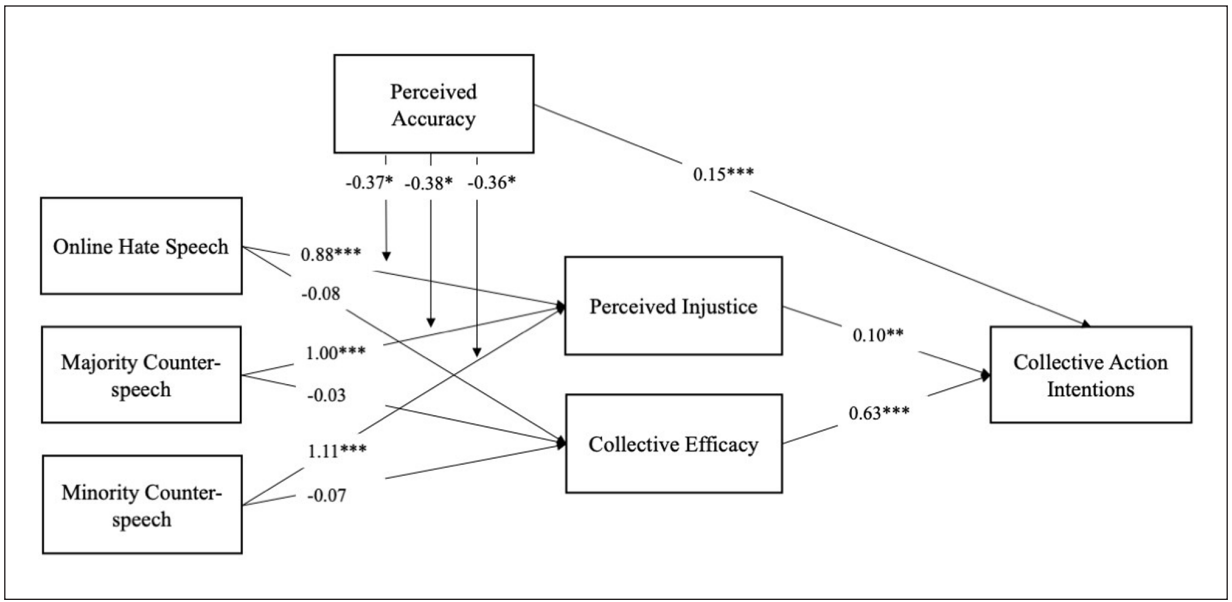


Figure 2. Results of path model in Study 1.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 2. Path Model Study 1 (Standalone HS is the Reference Group).

	Group-based injustice		Collective efficacy		Collective action	
Predictor	*b*	*SE*	*b*	*SE*	*b*	*SE*
Majority CS	.12	0.21	.04	0.15	−.13	0.14
Minority CS	.23	0.21	.01	0.14	.21	0.14
Control Group	−.88***	0.21	.08	0.15	.20	0.14
Group-based Injustice					.10**	0.04
Collective Efficacy					.63***	0.05
R^2	.10		.00		.33	

*Note.*HS = Hate Speech, CS = Counter-speech.
$N = 328$, $^{**}p < .01$. $^{***}p < .001$.

on collective action intentions via group-based injustice or collective efficacy had a 95% bootstrap confidence interval that excluded zero (see Table A1).

To investigate our RQ2, we recoded the variables again with majority counter-speech as reference group in the third path model. We found no significant difference between minority and majority counter-speech on group-based injustice ($b = .10$, $SE = 0.20$, $p = .606$) or efficacy ($b = -.04$, $SE = 0.14$, $p = .788$).

To test H3, we estimated a separate moderation model (see Table A0 in Supplemental Appendix A), which showed a very good model fit (CFI = 1.00, TLI = 1.00, RMSEA = .00, 90% CI [0.00, 0.09], SRMR = .01). Supporting H3a, perceived accuracy moderated the effects of all three treatment conditions on group-based injustice, such that the effects were stronger among participants who perceived the hate speech as less accurate (standalone hate speech: $b = -.37$, $SE = 0.18$, $p = .038$; majority counter-speech: $b = -.38$, $SE = 0.18$, $p = .029$; minority counter-speech: $b = -.36$, $SE = 0.17$, $p = .034$). In contrast, none of the interactions between perceived accuracy and the treatment conditions significantly predicted collective efficacy (standalone hate speech: $b = .11$, $SE = 0.11$, $p = .333$; majority counter-speech: $b = .04$, $SE = 0.12$, $p = .712$; minority counter-speech: $b = .13$, $SE = 0.11$, $p = .236$), not supporting H3b. The Johnson Neyman analysis indicated that those with perceived accuracy levels lower than 3.46 expressed higher feelings of group-based injustice after exposure to standalone hate speech, those with perceived accuracy levels lower than 3.66 for exposure to hate speech accompanied by minority counter-speech, and those with perceived accuracy levels lower than 3.35 for exposure to hate speech accompanied by majority counter-speech. No corresponding interaction effects emerged for collective efficacy; thus, H3b was not supported.

Discussion Study 1

Our findings show, in line with SIMCA, that group-based injustice and collective efficacy uniquely predict collective action intentions. Yet, online hate speech only

increased collective action via group-based injustice but not via collective efficacy. This lack of effect may be explained by the fact that online hate speech, in contrast to news coverage (Saleem et al., 2021), contains extreme insults and attacks that may lead to resignation or feelings of powerlessness for some individuals despite the activation and politicization of social identity. We also found no effect of counter-speech on group-based injustice or efficacy. It is possible that counter-speech did not exert any effect in the artificial experimental setting of the study. Another possible explanation is that we only included a single counter-speech comment, creating an equal ratio between hate speech and counter-speech. This balance may be insufficient to counteract the severe nature of the hate speech used in our study, as previous research suggests that counter-speech is more effective against milder forms of hate speech (see Van Houtven et al., 2024). To tackle this issue, we used a higher number of counter-speech comments in Study 2.

Another limitation of Study 1 is the fact that we only measured one specific type of collective action intention: engaging in collective action on social media, chosen for its proximity to the digital realm where hate speech occurs. However, collective action intentions vary, ranging from moderate, nonviolent actions to more extreme, radical behaviors. Literature commonly distinguishes between normative collective actions, such as peaceful protests or signing petitions, and non-normative actions, including sabotage, violent protests, or terrorism (Saleem et al., 2021; Tausch et al., 2011). While group-based injustice and collective efficacy may motivate moderate actions (van Zomeren, Postmes, et al., 2008), they may also increase the likelihood of non-normative responses (Wright et al., 1990). Therefore, we distinguish between normative and non-normative collective action intentions in Study 2 and look at collective action outside of the social media realm.

A further limitation of Study 1 is that perceived accuracy was measured at the level of the entire hate speech comment, making it unclear whether participants based their judgments on specific factual claims, the overall plausibility of the message, or its perceived social resonance. In addition, perceived accuracy of ingroup-directed hate speech exhibited moderate positive skewness (skew $= 0.55$, $p < .001$), reflecting that most participants perceived such content as inaccurate. Yet, the magnitude of the skew was modest, and linear models are generally robust to such deviations from normality, particularly with Likert-type measures and larger samples. To more directly assess the causal role of accuracy, Study 2 experimentally manipulates platform-flagged accuracy of the factual information embedded in the hate speech. This manipulation is not designed to evaluate real-world fact-checking practices, but to systematically vary the perceived legitimacy of the information used to justify hateful claims, thereby isolating the role of accuracy cues in shaping group-based injustice, collective efficacy, and collective action responses among targeted individuals.

Study 2

Study 2 extends Study 1 by (a) experimentally manipulating accuracy, (b) varying the number of counter-speech messages, and (c) distinguishing between normative and

non-normative collective action. Study 2 was designed as a conceptual extension of Study 1 rather than a direct replication. Although several elements were adapted to fit a different sociopolitical context—including the target group, the role of flagged accuracy, the inclusion of several counter-speech comments, and the distinction between normative and non-normative collective action—the two studies are unified by the same theoretical framework. Both examine how contextual cues surrounding online hate speech shape group-based injustice and collective action intentions, allowing Study 2 to test the generalizability of the mechanisms identified in Study 1 across contexts and action modalities.

Method

In Study 2, we used an online survey-experiment conducted from March until May 2023 with a sample of $N = 583$ US Latinos. We recruited participants through the survey company Prolific and the recruiting system of the University of California, Santa Barbara. The study was reviewed and approved by the ethical review board of the University of California, Santa Barbara (IRB ID: 7-23-0071). The study materials, dataset and analysis script are available on OSF: https://osf.io/kghjy/overview.

Design and Procedure. We conducted a between-subjects survey-experiment, manipulating the accuracy of the factual information in the hate speech and the type of counter-speech resulting in a 2 (accurate vs. inaccurate factual information) × 3 (no counter-speech, counter-speech by a minority member, or counter-speech by a majority member) design with a control group. We focused on Latino Americans, given the recent rise in xenophobic rhetoric and hate crimes targeting this growing demographic (The State of Hate Report, 2024). Upon assessing sociodemographic variables, we asked participants about their social media use, political ideology, their identification with their ethnic identity, and the representation of Latinos in American media. Participants were then randomly assigned to one of the seven conditions.

Inclusion criteria required participants to be 18 years or older and to be of Latino ethnicity. Since this target group is rather difficult to reach, participants for this study were recruited by means of a university recruitment tool and the survey company Prolific. The survey experiment concluded with language checks, including a section where participants estimated their comprehension of the survey items. This step aimed to ensure data quality control. As in Study 1, all participants provided informed consent, received a trigger warning in the beginning and a detailed debriefing at the end of the survey. We excluded nine participants, who requested the deletion of their data by checking the box at the end of the survey. Participants were incentivized with study credits or financial compensation (Prolific).

Respondents ($N = 583$ in the final sample) were randomly assigned to one of the seven experimental conditions: a control group ($n = 81$ or 14%), standalone hate speech with inaccurate information ($n = 82$ or 14%), hate speech with inaccurate information and minority counter-speech ($n = 84$ or 14%) or majority counter-speech ($n = 83$ or 14%), standalone hate speech with accurate information ($n = 83$ or 14%),

hate speech with accurate information and minority counter-speech ($n = 85$ or 15%) or majority counter-speech ($n = 85$ or 15%).

Sample. Of the 583 participants, 51% identified as female ($n = 299$), 46% as male ($n = 269$), and 3% of the participants identified as non-binary ($n = 15$). Participants' age ranged from 18 to 74 years ($M = 30.02$, $SD = 10.57$). With regard to education, 1.2% had no high school diploma, 42% had a high school degree, 19.4% an associate degree or vocational degree, 31% a Bachelor's degree, and 6.3% a postgraduate degree. The majority of participants reported being born in the United States ($n = 530$, 91%). Among those not born in the United States, the average amount of years lived in the United States was $M = 24.30$, $SD = 10.34$.

Stimulus Materials. Participants were exposed to a fictitious Facebook post shared by the fictional newspaper "The Independent," promoting an article titled "Latinos across the country embracing power in a shifting America." We created seven fictitious Facebook posts (see Supplemental Appendix B).

Analogous to Study 1, the online hate speech and counter-speech content drew upon prior research (Leonhard et al., 2018; Naab et al., 2018), with comments maintained at comparable lengths of 90 to 100 words. Again, a fictitious user "Alex Mei" was the author of the first comment. In the control condition, this comment was neutral stating "Interesting! Latinos in the US." In all six treatment conditions, the comment featured hate speech targeting Latinos including defamations and dehumanization ("wetbacks," "spiks"), expressing an exclusionary sentiment, and delegitimizing their presence in the United States ("We all agree that Latinos do not belong to America. GET LOST and go back where you came from! Americans first!").

In the counter-speech conditions, this comment was followed by three counter-speech comments. The counter-speech was represented by names typical for non-Latino majority members ("Hunter K.," "Allison," "Cody De") or other minority members ("Pablo K.," "Marianna," "Jose De"). We clarified the ethnic origin of the counter-speakers by including the following statements in the comments of the majority counter speakers "I AM NOT LATINO MYSELF," "I do not identify as Latina myself," "Latinos and other American citizens (like me)" and the following statements in the comments in the minority counter speakers "I AM LATINO MYSELF," "I do identify as Latina myself," "Latinos (like me)." The counter-speech countered the hate speech with arguments, stating ("These are just insinuations against Latinos in the United States, this is utterly inappropriate and wrong") and invoked shared values and unity in the same way as Study 1 (Dovidio et al., 2007).

We manipulated accuracy by using statistical numbers that are confirmed by the U.S. Department of Justice (Beck, 2021). The data suggested that of those arrested for (nonfatal) violent crime in 2018, 21% were Latino. To emphasize that this is correct statistical information, we added a fact-checking button stating "Approved by 3rd Fact Checker. Statistical information rated as true." To manipulate inaccuracy of the presented online hate speech, we used a considerably higher number of crimes committed

by Latinos (88%) and added a fact-checking label stating "Disputed by 3rd Party Fact-Checkers. Statistical information rated as false." The inaccurate statistic was selected based on a pretest (see Supplemental Appendix A). The label was modeled after a label previously used by Facebook to address hoaxes and misinformation (Mosseri, 2016). We chose this label to build upon prior research (Koch et al., 2023). However, because this exact label is no longer in use (Lyons, 2017), we presented it to participants as part of a beta-tested platform feature. This framing was intended to increase plausibility, as social media platforms have repeatedly introduced similar moderation and credibility features through pilot phases, such as Community Notes on Facebook, Instagram, and Threads (Meta Platforms, 2025).

We chose this combined manipulation (i.e., varying both the statistical information itself and the accompanying fact-checking label) to create a clear and controlled variation in both factual accuracy and participants' perceptions of credibility. Prior research shows that numerical claims can strongly shape political judgments, even when they are inaccurate (Stubenvoll & Matthes, 2022). At the same time, users are often unable to reliably assess factual accuracy on their own, and external cues such as fact-checking labels can shape credibility judgments (Koch et al., 2023). Given the limited research on the interplay between hate speech, misinformation, and counter-speech, we therefore opted for a strong and unambiguous manipulation. Our aim was not to replicate current platform moderation practices, but to isolate the role of factual accuracy and credibility cues under controlled conditions.

Measures. All constructs were measured on five-point Likert scales (1 = *strongly disagree*, 5 = *strongly agree*; Supplemental Appendix A presents an overview of all items, Table A4 the zero-order correlations). The survey was conducted in English. We measured *group-based injustice* of the hate speech (Cronbach's $\alpha = .95$, $M = 4.16$, $SD = 1.18$) based on Tausch et al. (2011) using four items. Items were introduced with the prompt "The next questions are related to Alex's comment. How do you evaluate it in general?" followed by "Alex's comment . . ." (1) "is unfair," (2) "is not justified," (3) "is unjust," and (4) "is not legitimate," which is in line with van Zomeren et al.'s (2018) and van Zomeren's (2020) conceptualization of perceived unfairness as one possible measurement of group-based injustice. As in Study 1, we gauged *collective efficacy* (Cronbach's $\alpha = .90$, $M = 4.25$, $SD = 0.76$) with three items introduced with "Please rate your agreement or disagreement with each of the statements below. Please be honest in your responses and remember your responses are anonymous" followed by (1) "Latinos in America can work together to fight against discrimination," (2) "Latinos in America can work together to improve their status and position in American society," (3) "Latinos in America can work together to exert influence in political decisions in America."

To gauge collective action, we first presented participants with the following prompt designed to elicit honest self-reports of behavioral intentions rather than abstract attitudes, while grounding responses in a realistic social and political context: "Sometimes people think that extreme actions are necessary or justified to change the ways in which their group is treated in the larger society. To what extent would YOU

be willing to engage in the following actions if they were part of a protest or demonstration aimed at changing the way Latinos are treated in American society. Please rate your agreement or disagreement with each of the statements below. Please be honest in your responses and remember your responses are anonymous.“ For *normative collective action intentions* (Cronbach's $\alpha = .87$, $M = 3.61$, $SD = 1.13$), we used four items by Saleem et al. (2021) (i.e., “I would. . .” (1) “participate in a demonstration with the goal of improving the position of Latinos in America,” (2) “participate in raising awareness about injustices facing Latinos in America,” (3) “sign a petition to stop discrimination against Latinos in America,” (4) “be willing to donate money to organizations that are intending to raise the concerns of Latino Americans.”) Participants indicated their *non-normative collective action intentions* (Cronbach's $\alpha = .90$, $M = 1.93$, $SD = 1.10$) by stating their agreement to three items by Saleem et al. (2021) and Tausch et al. (2011) (“I would . . .” (1) “disturb events,” (2) “block buildings,” and (3) “block the highway”). Finally, we again assessed whether participants perceived a threat to their social identity using five items, that is, “As Latino I feel. . .” (1) “discriminated,” (2) “degraded,” (3) “insulted,” (4)“offended,” (5) “threatened” “. . .by this content” ($\alpha = .94$, $M = 3.61$, $SD = 1.29$; Leonhard et al., 2018).

Power Analysis. We were interested in small main effects of (in)accurate information in online hate speech and the presence of (minority or majority) counter-speech (of around $f = .20$). Given an alpha-level of .05 and a power ($1 - \beta$) of 95%, to calculate an ANOVA with seven groups, we have thus aimed for at least 532 participants as our minimum sample size, which our final sample of $N = 583$ exceeded. A sensitivity analysis revealed that we were able to detect effects as small as $f = .19$.

Manipulation Check. We again found that participants perceived the group with which they identified to be more strongly disadvantaged in the hate speech condition ($M = 3.92$) than in the control condition, which was evaluated as largely neutral ($M = 1.71$), $t(111.91) = -18.597$, $p < .001$.

Furthermore, our counter-speech manipulation was successful. The association between the conditions and the perceived number of counter-speech comments was significant, $\chi^2(3, n = 583) = 516.41$, $p < .001$. Also, participants correctly estimated whether all counter-speakers were Latino $\chi^2(4, n = 583) = 234.96$, $p < .001$ and whether the counter-speech defended Latinos, $\chi^2(4, n = 583) = 53.05$, $p < .001$.

Moreover, participants in the treatment groups compared to the control group correctly recalled that the online hate speech included statistical information about crime committed by Latinos, $F(6, 571) = 38.87$, $p < .001$, $\eta^2_{part} = .29$. Along these lines, participants correctly indicated that the information in the online hate speech was “Disputed by 3rd Party Fact-Checkers,” $F(2, 580) = 263.7$, $p < .001$, $\eta^2_{part} = .47$, or “Approved by 3rd Party Fact-Checkers,” $F(2, 580) = 202.8$, $p < .001$, $\eta^2_{part} = .41$.

Randomization Check. Lastly, there were no systematic differences between conditions in terms of gender, $\chi^2(12) = 10.64$, $p = .56$, age, $F(6, 576) = 1.38$, $p = .222$, $\eta^2_{part} = .00$, education, $\chi^2(30) = 28.97$, $p = .519$, and social media usage, $F(6, 576) = .10$, $p = .997$, $\eta^2_{part} = .00$.

Data Analysis

To analyze the results, we ran three structural equation models (SEM) with latent variables for all mediators and dependent variable and manifest variables for the experimental conditions using the *lavaan* package in R with maximum likelihood estimation (Rosseel, 2012). To test H1, we conducted the first SEM with the full sample, including all seven experimental groups. For the second SEM, we excluded the control group, analyzing only the six treatment groups to test the remaining hypotheses and the research question (RQ). For the third SEM, we used again another set of dummy variables to compare the effects of majority and minority counter-speech. We used 5,000 bootstrapping samples to calculate 95% percentile confidence intervals for indirect effects.

Results

Our first SEM revealed a very good model fit (CFI = 0.98, TLI = 0.98, RMSEA = .04, 90% CI [0.03, 0.05], see Table A5 in Supplemental Appendix A). All treatment conditions, compared to the control group, significantly induced group-based injustice (standalone/low accuracy: $b = 2.74$, $SE = 0.15$, $p < .001$; minority/low accuracy: $b = 2.58$, $SE = 0.15$, $p < .001$; majority/low accuracy: $b = 2.65$, $SE = 0.15$, $p < .001$; standalone/high accuracy: $b = 2.63$, $SE = 0.15$, $p < .001$; minority/high accuracy: $b = 2.26$, $SE = 0.17$, $p < .001$; majority/high accuracy: $b = 2.50$, $SE = 0.15$, $p < .001$). In turn, group-based injustice predicted both normative ($b = .41$, $SE = 0.06$, $p < .001$) and non-normative collective action intentions ($b = .21$, $SE = 0.06$, $p < .001$).[1.]

For collective efficacy, we only found significant effects of the minority counter-speech/low accuracy ($b = .23$, $SE = 0.12$, $p = .049$) and the majority counter-speech/low accuracy ($b = .30$, $SE = 0.11$, $p = .008$) conditions in comparison to the control group, while all other effects were not statistically significant. Collective efficacy significantly influenced normative ($b = .62$, $SE = 0.08$, $p < .001$), but not non-normative collective action ($b = -.05$, $SE = 0.07$, $p = .439$). Hence, H1a was fully supported, while H1b was only partially supported (i.e., only for the conditions with low accuracy and counter-speech).

Additionally, significant indirect effects via group-based injustice were found for all treatment conditions on both normative and non-normative collective action intentions. Significant indirect effects via collective efficacy on normative collective action intentions were also found for the minority and majority counter-speech/low accuracy conditions (see Table A2).

Our second SEM (see Table 3, Figure 3) also revealed a very good fit (CFI = .97, TLI = .95, RMSEA = .06, 90% CI [0.05, 0.06]). Compared to the standalone hate speech condition, minority counter-speech significantly reduced group-based injustice ($b = -.26$, $SE = .08$, $p = .002$), whereas majority counter-speech had no such effect ($b = -.11$, $SE = .08$, $p = .204$), answering RQ1. To directly test RQ2 which explores whether the group identity of the counter speaker matters, we recoded the dummy variables to only compare the majority counter-speech and minority counter-speech

Table 3. Path Model Study 2 (Without Control Group).

	Group-based injustice		Collective efficacy		Normative collective action		Non-normative collective action	
Predictor	*b*	*SE*	*b*	*SE*	*b*	*SE*	*b*	*SE*
Majority CS	−.11	0.08	.07	0.08	.04	0.12	.14	0.12
Minority CS	−.26**	0.08	−.01	0.08	.04	0.12	.10	0.12
Accuracy	−.20**	0.07	−.12	0.07	−.00	0.09	.08	0.10
Group-based Injustice					.52***	0.07	.18*	0.07
Collective Efficacy					.60***	0.08	−.06	0.08
R^2	.04		.01		.27		.02	

Note. HS = Hate Speech, CS = Counter-speech.
$N = 502$, *$p < .05$. **$p < .01$. ***$p < .001$.

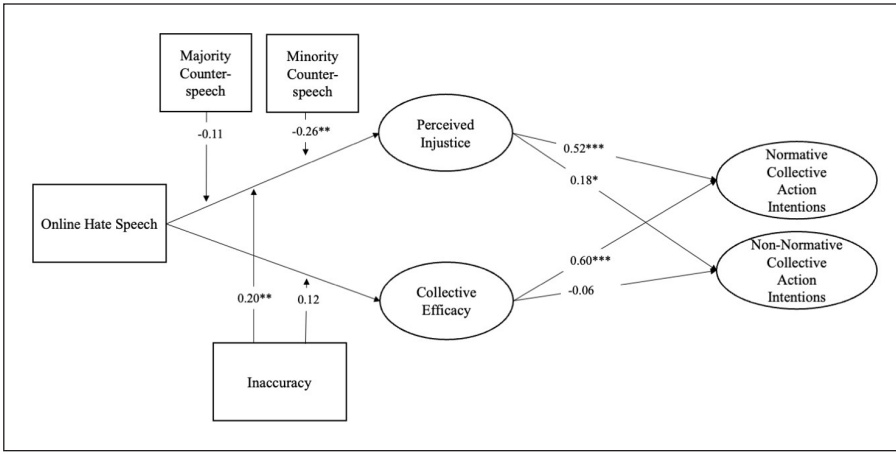


Figure 3. Results of structural equation model in Study 2 (without control group).
*$p < .05$. **$p < .01$. ***$p < .001$.

conditions in a third SEM. This difference was not statistically significant ($b = -.16$, $SE = 0.08$, $p = .056$).

Regarding collective efficacy, we did not find any significant effects of majority ($b = .07$, $SE = 0.08$, $p = .377$) or minority ($b = -.01$, $SE = 0.08$, $p = .920$) counter-speech compared to standalone hate speech. Likewise, the difference between majority and minority counter-speech ($b = -.08$, $SE = 0.08$, $p = .322$) was not statistically significant. Hence, H2 was not supported.

Furthermore, accuracy significantly decreased group-based injustice ($b = -.20$, $SE = 0.07$, $p = .004$), supporting H3a. For collective efficacy, this effect was not statistically significant ($b = -.12$, $SE = 0.07$, $p = .063$), not supporting H3b. Group-based injustice, in turn, was positively related to normative ($b = .52$, $SE = 0.07$, $p < .001$)

and non-normative ($b = .18$, $SE = 0.07$, $p = .014$) collective action, while collective efficacy was only related to normative ($b = .60$, $SE = 0.08$, $p < .001$), but not non-normative collective action intentions ($b = -.06$, $SE = 0.08$, $p = .436$).[2]

The indirect-effect analyses further showed that minority counter-speech, compared to standalone hate speech, indirectly reduced both normative and non-normative collective action intentions via lower group-based injustice. Greater accuracy likewise indirectly reduced both forms of collective action via lower group-based injustice. None of the corresponding indirect effects via collective efficacy had a 95% bootstrap confidence interval that excluded zero (see Table A2).

Discussion Study 2

Findings of Study 2 corroborate Study 1 in that the full path of online hate speech via group-based injustice on collective action was supported, as predicted by the SIMCA. This effect emerged for both normative and non-normative collective action intentions (Tausch et al., 2011). Regarding collective efficacy, we found higher efficacy compared to the control group only when low-accuracy hate speech was accompanied by either minority or majority counter-speech. By directly challenging the inaccuracies in hate speech, counter-speech may strengthen the group's belief in their capacity to act collectively and correct misrepresentations, thus enhancing collective efficacy (van Zomeren, Postmes, et al., 2008). However, because counter-speech did not significantly increase efficacy relative to standalone hate speech, this interpretation should be treated cautiously.

Interestingly, collective efficacy predicted normative but not non-normative collective action intentions. This pattern suggests that non-normative collective action is less driven by instrumental considerations regarding the feasibility or effectiveness of action, and more strongly rooted in appraisals of group-based injustice. In other words, when individuals perceive a situation as unfair, unjustified, or illegitimate, they may be motivated to endorse non-normative forms of action even in the absence of strong beliefs about collective efficacy (Saleem et al., 2021; van Zomeren et al., 2012). This distinction is crucial, as normative and non-normative collective actions carry different implications for legitimacy, public support, and the potential escalation of social change (Becker & Tausch, 2015).

As for counter-speech, our findings show that only minority counter-speech, but not majority counter-speech reduced group-based injustice compared to no counter-speech (and marginally also compared to majority counter-speech). Interestingly, counter-speech by a majority member had no such effect. As our counter-speech was focused on arguments promoting shared values and unity (Dovidio et al., 2007), one explanation for this effect may be that other minority members engaging in counter-speech serve as role models for constructive responses to hate speech, potentially reducing group-based injustice. Furthermore, the reason why we found an effect of counter-speech in Study 2 but not in Study 1 may be due to a higher number of counter-speech comments in Study 2 compared to only one counter-speech comment in Study 1.

Consistent with the pattern observed for perceived accuracy in Study 1, experimentally flagged accuracy in Study 2 decreased group-based injustice, which in turn was associated with lower normative and non-normative collective action intentions. System Justification Theory (Jost et al., 2017) explains this as individuals internalizing and justifying their disadvantaged status when accepting negative information as legitimate. This reduces the emotional drive for collective action, as the negative information is seen not as unjust but as an accepted reality, thereby diminishing the motivation to challenge it. Importantly, this does not imply that hate speech ceases to be harmful or emotionally provocative when accurate, but rather that factual legitimacy may attenuate group-based injustice that motivate collective action intentions, even when the message itself remains offensive or hostile. As such, flagged accuracy may reduce mobilization by reframing the situation as less unjustified, rather than by eliminating negative emotional responses altogether.

Finally, Study 2 shows remaining negative effects of all treatment conditions on normative and non-normative collective action intentions compared to the control group, when group-based injustice and collective efficacy are controlled. These findings suggest that exposure to hate speech may result in feelings of resignation and powerlessness that may even reduce the intention to act against the hate speech, when holding group-based injustice and collective efficacy constant.

General Discussion

Across two experiments involving minority groups in distinct national contexts (Muslims in Germany and Latinos in the United States), our findings demonstrate that exposure to online hate speech increases group-based injustice among targeted individuals. Consistent with the SIMCA (van Zomeren, Postmes, et al., 2008), this heightened group-based injustice, in turn, predicts greater collective action intentions. This finding aligns with prior research showing that identity-relevant threats and perceived injustice are key drivers of collective responses among disadvantaged groups (e.g., Branscombe et al., 1999; Major & O'Brien, 2005), and extends this work to the context of online hate speech. Importantly, these effects were shaped by communication-specific contextual cues. In both studies, group-based injustice was stronger when hateful claims were perceived as inaccurate or explicitly flagged as false, indicating that accuracy cues shape how targets appraise the fairness and legitimacy of group-based disadvantage. In this context, accuracy cues signal the societal legitimacy of group-based injustice. In other words, accuracy cues shape whether derogatory claims are interpreted as unjust distortions or as reflecting socially accepted realities. This finding complements prior research on misinformation and credibility cues, which shows that perceived accuracy influences how individuals evaluate and respond to contested information (e.g., Masullo et al., 2023; Pennycook et al., 2018), and extends it by demonstrating its relevance for identity-threatening content such as hate speech. Because inaccurate claims are common in online hate environments (Hameleers et al., 2022), these findings suggest that accuracy signals play an important role in shaping

collective-action dynamics. Whether subjectively perceived or embedded within digital platforms, such signals influence how identity-relevant threats are interpreted.

Consistent with past research, our findings highlight how different forms of collective action emerge from distinct psychological processes (van Zomeren et al., 2012). Normative collective action intentions were shaped jointly by group-based injustice and collective efficacy beliefs, whereas non-normative action intentions were more strongly associated with group-based injustice and less dependent on efficacy considerations (Tausch et al., 2011). This pattern is in line with prior work suggesting that non-normative actions are more strongly rooted in emotional and grievance-based responses, whereas normative actions depend more on instrumental considerations such as perceived efficacy (Tausch et al., 2011; van Zomeren et al., 2012). This distinction is consequential because non-normative responses, while sometimes strategically disruptive, may also escalate intergroup tensions depending on their form and context (Tausch et al., 2011). Notably, minority counter-speech reduced group-based injustice in Study 2, suggesting that ingroup opposition to hate speech can attenuate grievance responses that might otherwise motivate more confrontational responses. This finding contributes to the growing literature on counter-speech, which has primarily focused on attitudinal outcomes (e.g., Obermaier et al., 2023; Van Houtven et al., 2024), by demonstrating its potential to shape collective action processes among directly targeted individuals.

Limitations

Several limitations should be noted. First, we did not vary hate speech severity, though research suggests counter-speech may be more effective with milder hate (Van Houtven et al., 2024). While our study focused on severe hate, consistent with past research, future studies should compare different hate types for more detailed insights. Second, we measured intentions to engage in collective action rather than actual behavior, which is the most common way of measuring this construct (van Zomeren, Postmes, et al., 2008). Although intentions to engage in collective action are good predictors of actual participation (Moskalenko & McCauley, 2009), we do not know to what degree participants would actually perform these actions. Future research should try to capture actual participation—for instance, by providing participants with the possibility to sign a petition right after exposure to the media content. Third, although the fact-checking labels in our study were intended to manipulate perceived accuracy rather than to model real-world fact-checking interventions, and were directly modeled on labels previously used by Facebook (Mosseri, 2016), attaching such labels directly to user comments differs from how accuracy cues are more commonly implemented on social media platforms today and may therefore limit ecological validity.

At the same time, this design allowed us to create a clear manipulation of perceived accuracy in a context where participants could not be expected to know the correct statistical information themselves. Prior research shows that users are often unreliable in judging the accuracy of online information and that explicit fact-checking cues can shape credibility perceptions (Koch et al., 2023). Given the limited research on the interplay between hate speech, misinformation, and counter-speech, we therefore

opted for a strong and unambiguous first test. Future research should explore other practices, for example, by embedding accuracy cues directly into counter-speech or by using more naturalistic platform-based interventions such as community notes, visibility reduction, or warning labels. More broadly, future work should examine a wider range of community management strategies, from user flagging or reduced visibility to content removal or user bans (Pradel et al., 2024), to better understand how different interventions shape responses to online hate speech. This will become increasingly important as AI-generated content continues to proliferate on social media, making accuracy labels and related transparency cues more relevant for platform governance, as reflected in recent regulatory initiatives such as the European Commission's proposed use of content icons (European Commission, 2026).

Related to this point, the exposure situation does not fully mirror how individuals typically encounter such content in everyday online environments. These factors may have influenced the intensity of emotional responses, including anger, group-based injustice, or relief following counter-speech. At the same time, the controlled setting allowed us to systematically isolate the effects of flagged accuracy and counter-speech, increasing internal validity. Moreover, individuals who are frequently exposed to hate speech may be habituated to such content. Future research should therefore try to simulate more naturalistic contexts, such as field experiments or longitudinal exposure designs.

Fourth, the control condition contained less text than the treatments in both experiments. By providing only a brief, neutral comment, we aimed to minimize potential confounds. However, we cannot entirely rule out effects of text length on responses. Finally, although social identity was activated experimentally in the present studies, future research may benefit from directly measuring identity strength to assess how individual differences in identity activation interact with group-based injustice and efficacy in shaping collective action responses (see e.g., van Zomeren, Spears, et al., 2008).

Theoretical Implications

Our findings have important theoretical implications for understanding collective action in digital environments. The present research extends the Social Identity Model of Collective Action (SIMCA; van Zomeren, Postmes, et al., 2008) by identifying digital communication environments as an important source of the psychological appraisals that motivate collective action. Although SIMCA identifies group-based identity, injustice, and collective efficacy as central antecedents of collective action, it does not focus on if and how communication environments can shape the formation of these appraisals. The present findings suggest that digital communication environments are not merely contexts in which collective action unfolds but actively contribute to the formation of the psychological processes specified by SIMCA.

More specifically, the present research extends earlier work demonstrating that perceived accuracy shapes responses to negative media portrayals (Saleem et al., 2021) by showing that similar legitimacy processes operate in the context of online hate speech. Even when messages contain explicit hostility, communication cues indicating whether derogatory claims are legitimate or illegitimate shape subsequent injustice

appraisals. This shifts attention from hostility alone to the broader communication context in which hostility is encountered, suggesting that communicative features of online environments constitute important antecedents of injustice appraisals.

The findings also provide new insight into the development of collective efficacy within digital environments. Rather than viewing efficacy as arising solely from shared group membership or direct experiences of disadvantage, the present findings suggest that efficacy beliefs may also be shaped through observing communicative expressions (e.g., Velasquez & Quenette, 2018) that others are willing to challenge injustice. Such cues may represent one mechanism through which efficacy beliefs develop in digitally mediated environments. Collectively, these findings suggest that communication environments should be understood not simply as contexts in which collective action occurs, but as environments that actively shape the legitimacy and efficacy appraisals through which collective action emerges.

Practical Implications

The findings also carry practical implications for platform governance and user responses to online hate speech. Fact-checking labels that clearly signal the inaccuracy of hateful claims can influence how targets interpret these messages and may reduce concerns pertaining to the legitimization of this information. At the same time, our findings suggest that flagging such claims as inaccurate can heighten perceptions of group-based injustice, likely because the derogatory claims are perceived as less legitimate or justified. Moreover, other harmful elements of hate speech—such as derogatory language—can still provoke negative responses like outgroup avoidance (Van Houtven et al., 2024) or even hateful reactions (Obermaier et al., 2023). Addressing both the informational and expressive elements of hate speech, through visible fact-checking and/or counter-speech, may therefore be necessary to limit harmful downstream consequences for targeted groups.

Our findings further suggest that multiple visible counter-speech messages may be more effective than isolated responses. Future research should test this possibility directly. Encouraging visible counter-speech by multiple users may therefore represent a promising strategy for mitigating the psychological and behavioral consequences of online hate environments.

ORCID iDs

Desiree Schmuck https://orcid.org/0000-0002-9492-6052
Magdalena Obermaier https://orcid.org/0000-0002-3055-3744
Nancy Molina-Rogers https://orcid.org/0000-0003-4131-6186
Muniba Saleem https://orcid.org/0000-0002-6610-9880

Ethical Considerations

This research was approved by the Internal Review Board of the Department of Communication of UC Santa Barbara (ID:7-22-0596).

Consent to Participate

Written informed consent was obtained from each participant.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

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

Open Practices



The data is available at the open repository OSF: The study materials, dataset and analysis script are available on OSF: https://osf.io/kghjy/overview

Supplemental Material

Supplemental material for this article is available online.

Notes

1. As a robustness check we also ran the same model with group-based injustice operationalized as affective anger (consistent with Study 1, for exact items see Supplemental Appendix). Findings of this analysis were consistent with Study 1 with group-based injustice operationalized as affective anger significantly predicting normative ($b = .48$, $SE = .05$, $p < .001$) and non-normative ($b = .33$, $SE = .05$, $p < .001$) collective action, while collective efficacy only predicted normative ($b = .59$, $SE = 0.07$, $p < .001$), but not non-normative collective action ($b = -.10$, $SE = 0.07$, $p = .126$), supporting the interpretation that evaluative injustice appraisals and anger-based responses reflect complementary expressions of the same underlying group-based injustice experience (van Zomeren, 2020; van Zomeren et al., 2018).
2. In Study 2, we also assessed collective action with the same three items as in Study 1 (Cronbach's $\alpha = .82$, $M = 3.32$, $SD = 1.08$, see items in Supplemental Appendix A). When using this scale as dependent variable, we directly replicated the pattern from Study 1, showing that both group-based injustice ($b = .21$, $SE = .03$, $p < .001$) and collective efficacy ($b = .12$, $SE = .03$, $p < .001$) predicted collective action intentions on social media.

References

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect:" Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, *19*(3), 373–387. https://doi.org/10.1111/jcc4.12009

Beck, A. J. (2021). *Race and ethnicity of violent crime offenders and arrestees, 2018, Statistical Brief*. https://bjs.ojp.gov/content/pub/pdf/revcoa18.pdf

Becker, J. C., & Tausch, N. (2015). A dynamic model of engagement in normative and non-normative collective action: Psychological antecedents, consequences, and barriers. *European Review of Social Psychology*, *26*(1), 43–92. https://doi.org/10.1080/10463283.2015.1094265

Bettencourt, B. A., Charlton, K., Dorr, N., & Hume, D. L. (2001). Status differences and ingroup bias: A meta-analytic examination of the effects of status stability, status legitimacy, and group permeability. *Psychological Bulletin*, *127*(4), 520–542. https://doi.org/10.1037/0033-2909.127.4.520

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Boeckmann, R. J., & Liew, J. (2002). Hate speech: Asian American students' justice judgments and psychological responses. *Journal of Social Issues*, *58*(2), 363–381. https://doi.org/10.1111/1540-4560.00265

Branscombe, N. R., Schmitt, M. T., & Harvey, R. D. (1999). Perceiving pervasive discrimination among African Americans: Implications for group identification and well-being. *Journal of Personality and Social Psychology*, *77*(1), 135–149. https://doi.org/10.1037/0022-3514.77.1.135

Buerger, C., & Wright, L. (2019). Counterspeech: A literature review. *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.3829816

Delgado, R., & Stefancic, J. (2019). *Understanding words that wound*. Routledge. https://doi.org/10.4324/9780429503351

Dovidio, J. F., Gaertner, S. L., & Saguy, T. (2007). Another view of "we": Majority and minority group perspectives on a common ingroup identity. *European Review of Social Psychology*, *18*(1), 296–330. https://doi.org/10.1080/10463280701726132

Dredge, R., Gleeson, J. F. M., & de la Piedad Garcia, X. (2014). Risk factors associated with impact severity of cyberbullying victimization: A qualitative study of adolescent online social networking. *Cyberpsychology, Behavior, and Social Networking*, *17*(5), 287–291. https://doi.org/10.1089/cyber.2013.0541

Drury, J., & Reicher, S. (2009). Collective psychological empowerment as a model of social change: Researching crowds and power. *Journal of Social Issues*, *65*(4), 707–725. https://doi.org/10.1111/j.1540-4560.2009.01622.x

European Commission. (2026, June 10). EU Icons for Labelling AI-Generated Content. https://digital-strategy.ec.europa.eu/en/policies/eu-icons-labelling-ai-generated-content

Glasford, D. E., & Calcagno, J. (2012). The conflict of harmony: Intergroup contact, commonality and political solidarity between minority groups. *Journal of Experimental Social Psychology*, *48*(1), 323–328. https://doi.org/10.1016/j.jesp.2011.10.001

Hameleers, M., Van Der Meer, T., & Vliegenthart, R. (2022). Civilized truths, hateful lies? Incivility and hate speech in false information – evidence from fact-checked statements in the US. *Information, Communication & Society*, *25*(11), 1596–1613. https://doi.org/10.1080/1369118x.2021.1874038

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, *38*(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave Your Comment Below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research*, *41*(4), 557–576. https://doi.org/10.1111/hcre.12059

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, *33*(1), 1–27. https://doi.org/10.1111/j.2044-8309.1994.tb01008.x

Jost, J. T., Becker, J., Osborne, D., & Badaan, V. (2017). Missing in (collective) action: Ideology, system justification, and the motivational antecedents of two types of protest behavior. *Current Directions in Psychological Science*, *26*(2), 99–108. https://doi.org/10.1177/0963721417690633

Kim, Y. (2024). Shame on you! How incivility and absence of supporting evidence in like-minded Facebook comments influence evaluations of ingroup members and online political participation. *Online Information Review*, *48*(3), 619–643. https://doi.org/10.1108/oir-12-2021-0650

Koch, T. K., Frischlich, L., & Lermer, E. (2023). Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media. *Journal of Applied Social Psychology*, *53*(6), 495–507. https://doi.org/10.1111/jasp.12959

Leader Maynard, J., Benesch, S., & The University of Oxford. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, *9*(3), 70–95. https://doi.org/10.5038/1911-9933.9.3.1317

Leets, L. (2002). Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of Social Issues*, *58*(2), 341–361. https://doi.org/10.1111/1540-4560.00264

Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *Studies in Communication | Media*, *7*(4), 555–579. https://doi.org/10.5771/2192-4007-2018-4-555

LfM. (2024). *Hate Speech Forsa-Studie 2023 . Zentrale Untersuchungsergebnisse*. https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Hass/forsa_LFMNRW_Hassrede2023_Praesentation.pdf

Lyons, T. (2017, December 20). Replacing Disputed Flags with Related Articles. https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/

Major, B., & O'Brien, L. T. (2005). The social psychology of stigma. *Annual Review of Psychology*, *56*(1), 393–421. https://doi.org/10.1146/annurev.psych.56.091103.070137

Masullo, G. M., Tenenboim, O., & Lu, S. (2023). "Toxic atmosphere effect": Uncivil online comments cue negative audience perceptions of news outlet credibility. *Journalism*, *24*(1), 101–119. https://doi.org/10.1177/14648849211064001

McDonald, P., Charlesworth, S., & Graham, T. (2016). Action or inaction: Bystander intervention in workplace sexual harassment. *The International Journal of Human Resource Management*, *27*(5), 548–566. https://doi.org/10.1080/09585192.2015.1023331

Meta Platforms. (2025, March 13). Testing Begins for Community Notes on Facebook, Instagram and Threads. https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/

Moskalenko, S., & McCauley, C. (2009). Measuring political mobilization: The distinction between activism and radicalism. *Terrorism and Political Violence*, *21*(2), 239–260. https://doi.org/10.1080/09546550902765508

Mosseri, A. (2016, December 15). News Feed FYI: Addressing Hoaxes and Fake News. About Meta. https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/

Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, *20*(2), 777–795. https://doi.org/10.1177/1461444816670923

Obermaier, M., Schmuck, D., & Saleem, M. (2023). I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *New Media & Society*, *25*(9), 2339–2358. https://doi.org/10.1177/14614448211017527

Outten, H. R. N., & Schmitt, M. T. (2015). The more "intergroup" the merrier? The relationship between ethnic identification, coping options, and life satisfaction among South Asian Canadians. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *47*(1), 12–20. https://doi.org/10.1037/a0036871

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, *9*(1), 171–180. https://doi.org/10.17645/mac.v9i1.3399

Pabian, S. (2025). Bystanders' perceptions on online hate speech: Investigating the effects of perpetrators' justifications and the bystander's role on bystanders' attitude and prosocial intervention intentions. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *19*(4), 6. https://doi.org/10.5817/cp2025-4-6

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pradel, F., Zilinsky, J., Kosmidis, S., & Theocharis, Y. (2024). Toxic speech and limited demand for content moderation on social media. *American Political Science Review*, *118*(4), 1895–1912. https://doi.org/10.1017/s000305542300134x

Rosseel, Y. (2012). lavaan: AnRPackage for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

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

Saleem, M., & Ramasubramanian, S. (2019). Muslim Americans' responses to social identity threats: Effects of media representations and experiences of discrimination. *Media Psychology*, *22*(3), 373–393. https://doi.org/10.1080/15213269.2017.1302345

Saleem, M., Wojcieszak, M. E., Hawkins, I., Li, M., & Ramasubramanian, S. (2019). Social identity threats: How media and discrimination affect Muslim Americans' identification as Americans and trust in the U.S. government. *Journal of Communication*, *69*(2), 214–236. https://doi.org/10.1093/joc/jqz001

Schäfer, S., Rebasso, I., Boyer, M. M., & Planitzer, A. M. (2024). Can we counteract hate? Effects of online hate speech and counter speech on the perception of social groups. *Communication Research*, *51*(5), 553–579. https://doi.org/10.1177/00936502231201091

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2024). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, *26*(5), 2614–2632. https://doi.org/10.1177/14614448221091185

Shmulewitz, D., Levitin, M. D., Skvirsky, V., Vider, M., Lev-Ran, S., & Mikulincer, M. (2025). Exposure to online hate speech is positively associated with post-traumatic stress disorder symptom severity. *Scientific Reports*, *15*(1), Article 29869. https://doi.org/10.1038/s41598-025-16168-1

Stahel, L., & Baier, D. (2023). Digital hate speech experiences across age groups and their impact on well-being: A nationally representative survey in Switzerland. *Cyberpsychology, Behavior, and Social Networking*, *26*(7), 519–526. https://doi.org/10.1089/cyber.2022.0185

Stubenvoll, M., & Matthes, J. (2022). Why retractions of numerical misinformation fail: The anchoring effect of inaccurate numbers in the news. *Journalism & Mass Communication Quarterly*, *99*(2), 368–389. https://doi.org/10.1177/10776990211021800

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Nelson-Hall.

Tausch, N., Becker, J. C., Spears, R., Christ, O., Saab, R., Singh, P., & Siddiqui, R. N. (2011). Explaining radical group behavior: Developing emotion and efficacy routes to normative and nonnormative collective action. *Journal of Personality and Social Psychology*, *101*(1), 129–148. https://doi.org/10.1037/a0022728

The State of Hate Report. (2024). The Leadership Conference Education Fund. https://civilrights.org/edfund/wp-content/uploads/sites/2/2024/05/CauseforConcern-TheStateofHate-2024.pdf

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on twitter. *Sage Open*, *10*(2), Article 2158244020919447. https://doi.org/10.1177/2158244020919447

Van Houtven, E., Acquah, S. B., Obermaier, M., Saleem, M., & Schmuck, D. (2024). 'You got my back?' Severity and counter-speech in online hate speech toward minority groups. *Media Psychology*, *27*(6), 923–954. https://doi.org/10.1080/15213269.2023.2298684

van Zomeren, M. (2020). Political and psychological processes in political action. In E. Hannah (Ed.), *Oxford Research Encyclopedia of Politics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1030

van Zomeren, M., Kutlaca, M., & Turner-Zwinkels, F. (2018). Integrating who "we" are with what "we" (will not) stand for: A further extension of the Social Identity Model of Collective Action. *European Review of Social Psychology*, *29*(1), 122–160. https://doi.org/10.1080/10463283.2018.1479347

van Zomeren, M., Leach, C. W., & Spears, R. (2012). Protesters as "Passionate Economists." *Personality and Social Psychology Review*, *16*(2), 180–199. https://doi.org/10.1177/1088868311430835

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., Spears, R., & Leach, C. W. (2008). Exploring psychological mechanisms of collective action: Does relevance of group identity influence how people cope with collective disadvantage? *British Journal of Social Psychology*, *47*(2), 353–372. https://doi.org/10.1348/014466607x231091

Van't Riet, J., & Van Stekelenburg, A. (2022). The effects of political incivility on political trust and political participation: A meta-analysis of experimental research. *Human Communication Research*, *48*(2), 203–229. https://doi.org/10.1093/hcr/hqab022

Velasquez, A., & Quenette, A. M. (2018). Facilitating social media and offline political engagement during electoral cycles: Using social cognitive theory to explain political action among Hispanics and Latinos. *Mass Communication and Society*, *21*(6), 763–784. https://doi.org/10.1080/15205436.2018.1484489

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In S. O. Sood, M. Nagarajan, & M. Gamon (Eds.), *Proceedings of the second workshop on language in social media* (pp. 19–26). Association for Computational Linguistics.

Wright, S. C., Taylor, D. M., & Moghaddam, F. M. (1990). Responding to membership in a disadvantaged group: From acceptance to collective protest. *Journal of Personality and Social Psychology*, *58*(6), 994–1003. https://doi.org/10.1037/0022-3514.58.6.994

Ziegele, M., Naab, T. K., & Jost, P. (2020). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society*, *22*(5), 731–751. https://doi.org/10.1177/1461444819870130

Author Biographies

Desiree Schmuck (PhD University of Vienna) is a Professor of Communication Science in the Department of Communication of the University of Vienna, where she directs the Vienna Media Change and Media Innovation Group (VMCI). Her research focuses on changes in the media environment due to the presence of digital technologies and artificial intelligence (AI) and their consequences for media users.

Magdalena Obermaier (PhD LMU Munich), is a postdoctoral researcher in the Department of Media and Communication at LMU Munich, Germany. Her research focuses on (digital) media effects, specifically effects of pro- and antisocial communication, effects of persuasive communication, and the relationship between journalism and public relations.

Nancy Molina-Rogers (PhD University of California, Santa Barbara) is a postdoctoral fellow in the Department of Communication at the University of California, Los Angeles. Her research focuses on the relationship between media and political participation among marginalized groups, with a particular emphasis on racial and ethnic identity.

Muniba Saleem (PhD Iowa State University) is a professor in the Department of Communication at the University of California, Santa Barbara and a faculty associate within the Institute for Social Research at University of Michigan. She studies how media affect interpersonal and intergroup relations between racial, ethnic, and religious groups using social scientific methods.