

# Do “Good Citizens” fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments

Marlene Kunst, Pablo Porten-Cheé, Martin Emmer & Christiane Eilders

**To cite this article:** Marlene Kunst, Pablo Porten-Cheé, Martin Emmer & Christiane Eilders (2021) Do “Good Citizens” fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments, Journal of Information Technology & Politics, 18:3, 258-273, DOI: [10.1080/19331681.2020.1871149](https://doi.org/10.1080/19331681.2020.1871149)

**To link to this article:** <https://doi.org/10.1080/19331681.2020.1871149>



© 2021 The Author(s). Published with  
license by Taylor & Francis Group, LLC.



Published online: 11 Jan 2021.



Submit your article to this journal [↗](#)



Article views: 3730



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 14 View citing articles [↗](#)

## Do “Good Citizens” fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments

Marlene Kunst , Pablo Porten-Cheé , Martin Emmer , and Christiane Eilders

### ABSTRACT

In an effort to counter hate speech, media platforms have increasingly come to rely on ordinary users to fight abusive content. However, little is known about the predictors of this type of user engagement, which we refer to as online civic intervention (OCI). This article presents an experimental inquiry (N = 337) into whether solidarity citizenship norms promote OCI. The results show that users who support solidarity citizenship norms tend to have a greater propensity to flag hate comments and to engage in counterspeech. Overall, this indicates that “good citizens” are more inclined to stand up against hate speech online.

### KEYWORDS

Counterspeech; hate speech; user comments; flagging; citizenship norms; political participation

The dissemination of hate speech in communicative online spaces represents one of the major societal challenges of digitalization. Hate speech – defined as abusive expressions that incite violence, hatred, or discrimination of people on the basis of their belonging to a social group (Erjavec & Kovacic, 2012) – constitutes a severe threat to the deliberative potential of online spaces. It does not only disrupt meaningful issue-related political discussions, but may also discourage individuals and vulnerable social groups from participating in online debates and voicing their opinion (Munger, 2017). Moreover, hate speech can cause psychological damage to those who are verbally attacked (Boeckmann & Liew, 2002; Delgado & Stefancic, 2019). In search of solutions, online platforms – e.g., news sites or social media companies – are resisting harmful user-generated content not only by employing professional content moderators or applying artificial intelligence but also by relying on users to intervene when they are exposed to abusive language (Gillespie, 2018). Thus, ordinary users play a pivotal role in safeguarding the public discourse and reducing the hateful content that circulates online (Munger, 2017). However, so far, little is known about the predictors of this type of user engagement, which we refer to as online civic intervention (OCI) (Porten-Cheé, Kunst, & Emmer, 2020).

Essentially, engagement in OCI is likely to be predicted by both the content characteristics of a comment and the individual characteristics of the user, who is exposed to it. The present study sets out to test effects of both of these factors. Firstly, to find further support for previous empirical findings (Kalch & Naab, 2017; Leonhard et al., 2018; Wilhelm, Joeckel, & Ziegler, 2019), we investigate whether the willingness to engage in OCI increases when users are exposed to comments that contain obvious hate speech compared to more subtle disparaging forms of speech. Secondly, we examine the impact of an individual's support for citizenship norms on engagement in OCI. Thereby, we aim to add to the extant knowledge of a small but growing body of research that has shown that individual characteristics, such as personal attitudes toward social groups on the receiving end of hate speech (Kalch & Naab, 2017) or individuals' moral orientation (Wilhelm & Joeckel, 2018; Wilhelm et al., 2019), influence engagement in OCI.

Drawing on a range of studies that have established a positive association between citizenship norms and political participation (e.g., Copeland, 2014; Dalton, 2006; Theiss-Morse, 1993), we essentially assume that individuals with strong norm conceptions of what it takes to be a good citizen would feel more obliged to engage in OCI. To

validate these hypotheses, we present a web-based survey experiment in which one experimental group was exposed to a hate comment, while the other experimental group was exposed to a comment that disparages the same social group but lacks abusive language. The present study examined user responses to comments beneath news articles because comment sections have been severely affected by hate speech (Erjavec & Kovacic, 2012; Gardiner et al., 2016). Moreover, news article comment sections attract many readers and are usually not algorithmically customized to individual audience members. Thus, unlike in social media environments, exposure to hate speech in comment sections is not dependent on a user's network or previous behavior. Even individuals who are very unlikely to be exposed to hate speech in social media environments may be exposed to hate comments in comment sections, which increases the external validity of our experimental study.

### Online civic intervention in online discussions

Reporting abusive content to online platform providers, engaging in counterspeech, and rating user comments through social buttons constitute part of a wider spectrum of user activities that we refer to as OCI. We define OCI as action taken by ordinary users to fight disruptive online behavior with the aim of restoring civil and rational public discourse (Porten-Che   et al., 2020). Moreover, we argue that OCI is a new type of political participation that takes place in the digital public sphere. Following the conceptual route of Theocharis and van Deth (2017), a phenomenon qualifies as political participation if it either pursues the goal of changing political outcomes and/or refers to a political context. Although the aim of OCI may not be to directly influence political decisions, we argue that it qualifies as political participation because it seeks to create proper conditions for inclusive political discussions. Victims of hate speech may, for instance, withdraw from political discussions online but re-integrate if they feel socially supported by other users who come to their defense. Only if the political public discourse is inclusive and accessible to all, public opinion formation can take into account the heterogeneous

viewpoints and diverse voices that exist in contemporary societies.

### Different types of OCI

Essentially, we distinguish between two forms of OCI: low-threshold OCI, which is enabled by media platform tools (e.g., social buttons flagging), and high-threshold OCI, which is a verbal and discursive form of user involvement commonly referred to as counterspeech (Porten-Che   et al., 2020). High-threshold OCI takes more effort than low-threshold OCI and is more publicly visible. When exposed to hate speech, high-threshold OCI can, for instance, mean that individuals enjoin the commenter to be respectful toward others. An example of high-threshold OCI in the case of hate speech is a hashtag activist group known as *#jag  rh  r* [*#iamhere*] in Sweden and *#ichbinhier* in Germany, which attempts to collectively fight hate speech through polite counterspeech (Ley, 2018).

Empirical findings have indicated that users tend to engage in a single type of OCI (Kalch & Naab, 2017), but little is known about why users choose one type over the other. Kalch and Naab (2017) argued that users tend to prefer flagging over counterspeech because the latter requires more effort. Moreover, users may believe that, compared to pushing buttons, engaging in counterspeech requires many personal skills, such as writing literacy, knowledge of the issue in question, and internal political self-efficacy (Jost, Ziegele, & Naab, 2020). Thus, individuals who believe that they lack these skills may be more comfortable pushing social buttons or flagging than employing counterspeech.

Yet, Kalch and Naab (2017) found that users are more likely to flag than dislike strongly deviant content. Kalch and Naab (2017) argued that flagging may be perceived as more efficient than disliking because it alerts professional moderators, who may subsequently choose to block the commenter or delete the content. Nevertheless, rating user comments through social buttons is an integral part of the action repertoire of users who aim to fight hate speech (Jost et al., 2020). After all, social buttons give users the opportunity to publicly endorse or condemn other users' comments while

flagging takes place behind the scenes (Porten-Cheé et al., 2020; Knobloch-Westerwick, Sharma, Hansen, & Alter, 2005). In this light, Watson, Peng, and Lewis (2019) conceptualized the response to deviant content through social buttons as a more direct type of social control than flagging as social buttons allow users to send immediate public signals to the commenter and the community. Social control is defined as “a series of interventions that encourage pro-social behaviors and discourage deviant behaviors” (Watson et al., 2019, p. 1853). The authors emphasize that users need to engage in direct and indirect types of social control to tackle such a complex phenomenon as hate speech.

Lastly, the different types of OCI vary in terms of social consequences. When engaging in counter-speech, the comments are visible to other users, who may react to them (Jost et al., 2020; Ley, 2018). Therefore, conflict-avoidant individuals may prefer flagging or social buttons over counter-speech out of fear that the latter may entangle them in a dispute with other users (Jost et al., 2020). Yet, since many social platforms also show which users have pushed a social button, this activity can be public as well and make users feel more exposed than flagging. While the public nature of counter-speech and, to some extent, social buttons may trigger discomfort among some users, it may encourage those who expect social recognition from their engagement in OCI (Jost et al., 2020).

Although some users may prefer one type of OCI over the other, pushing social buttons, flagging content, and engaging in counterspeech are ways in which ordinary users aim to fight hate speech. The following section will elaborate on the content characteristics of user comments that make all types of OCI especially likely.

### **Hate speech vs. subtle forms of disparagement**

The most obvious versions of hate speech contain defamations, insults, or incitements to violence (Chen, 2017; Erjavec & Kovacic, 2012; Hanzelka & Schmidt, 2017; Wilhelm et al., 2019). However, hate speech can also be more subtle, and the lines between solely disparaging acts of speech and hate speech may be blurry at times. The strength of the language in a user comment and its severity in

terms of incitement to violence will affect not only the user's criminal liability but also other users' responses. Thus, a range of studies have found that incitements to violence and strongly abusive language tend to be reported more often than more subtle forms of disparagement (Kalch & Naab, 2017; Leonhard et al., 2018; Wilhelm et al., 2019). According to Leonard et al. (2018), this effect can be explained using the bystander theory (see Latané & Darley, 1970), which essentially claims that the more an incident appears to be an emergency, the more likely individuals are to help the victims. Translated to online environments, the more threatening and harmful the act of speech, the more alarming the situation will appear to other users, which will, in turn, increase the likelihood that they will intervene (Leonard et al., 2018). Similarly, scholars have argued that the likelihood of users engaging in OCI is dependent on the content's degree of social deviance; hence, the more strongly commenters violate social norms, the more resistance they can expect from other users who aim to restore social order (Wilhelm et al., 2019).

Drawing on the social maintenance model of retributive justice, Boeckmann and Liew (2002, p. 367) claimed that individuals tend to demand more severe punishments for deviant behaviors that are perceived as violations against norms “that are essential to group functioning” and may, thus, have a strong social impact. Compared to subtle forms of disparagement, hate speech is more likely to be perceived as a threat to the attacked social group due to its abusive and violence-invoking language. Consequently, as users are likely to demand more severe punishments for such harmful content, they may also be more likely to engage in OCI as it may feel like a way to contribute to the commenter's punishment. Lastly, users may also distinguish between views that are protected by laws, such as laws that guarantee freedom of expression in Germany, and views that are not protected. In the case of obvious hate speech, there are certain cues, such as strongly abusive language or incitements to violence, that indicate that content is likely to be prohibited by law (Sirsch, 2013). As more subtle forms of disparagement lack such cues, users may assume that this content must be tolerated due to freedom of expression and, thus, refrain from intervening.

Overall, the rationales set out above indicate that the more severe and obvious the attacks on a social group, the higher the likelihood that the other users will intervene. Consequently, we assumed that individuals are generally more likely to respond to user comments containing hate language with counter-speech, social buttons, or flagging than to user comments comprising more subtle, disparaging language. Although the types of social buttons vary, the majority of social media and news platforms provide users with an opportunity to rate other users' content (Singer, 2014). Therefore, in the present study, we tested participants' willingness to use the common and well-known dislike thumb button.

Against this backdrop, we posited the following hypotheses:

H1–3: Participants will be more willing to dislike (H1), flag (H2), and engage in counterspeech (H3) when exposed to a user comment that contains hate speech than a user comment that contains disparaging speech.

### Citizenship norms and OCI

Research has shown that individuals' characteristics, such as moral foundations (Wilhelm & Joeckel, 2018; Wilhelm et al., 2019) or trust in media (Watson et al., 2019), tend to influence whether they engage in OCI. We aim to extend this knowledge by suggesting that the willingness to engage in OCI is likely to be particularly strong for individuals who believe that it is a citizen's responsibility to stand up for others. Thus, individuals' attitudes about what it takes to be a good citizen may be a pivotal predictor of OCI. In academic research, expectations toward citizenship are commonly conceptualized as citizenship norms (Dalton, 2008; van Deth, 2007). Dalton (2008) suggested that this set of expectations can be divided into so-called dutiful and engaged citizenship norms. Dutiful citizenship norms comprise expectations about individuals' relations with government and actions that contribute to public obedience and social order. Engaged citizenship norms emphasize an individual's social responsibility and "willingness to act on his or her principles, be politically independent and address social needs" (Dalton, 2008, p. 81).

So far, empirical research has uncovered associations between dutiful citizenship norms and traditional institutionalized political participation (e.g., voting or joining a political party) and between engaged citizenship norms (also referred to as self-actualizing citizenship norms; Bennett, 2008) and types of participation that do not directly aim to influence political decision-making, such as community volunteering or consumerism (Chang, 2016; Copeland, 2014). Moreover, research has been concerned with how digital media use is related to citizenship norms and how this may, in turn, affect political participation (Feezell, Conroy, & Guerrero, 2016; Ohme, 2018; Thorson, 2015). For instance, Feezell et al. (2016) found that online behavior associated with dutiful citizenship (e.g., looking for the positions of political candidates) had a more positive effect on traditional political engagement than online behavior associated with engaged citizenship (e.g., discussing about political issues with others). However, not all studies identified clear distinctions between dutiful and engaged citizenship norms (Hooghe, Oser, & Marien, 2016), and empirical research has challenged the clear association between respective citizenship norms and types of political participation (Bolzendahl & Coffé, 2013; Copeland & Feezell, 2017).

Generally, OCI can be regarded as a new form of political participation that aims to contribute to a civil and rational discourse in the digital public sphere. Thus, it must be distinguished from conventional types of participation that mostly take place in the offline space or aim to directly influence political decision-making. With regard to hate speech, users who engage in OCI come to the defense of vulnerable social groups. For this reason, we do not distinguish between dutiful or engaged citizenship norms as potential predictors of OCI; but instead focus on citizenship norms that are associated with a sense of solidarity with others. In this study, we labeled these norms as solidarity citizenship norms because they imply the belief that good citizens are expected to support and care for others.

Individuals who feel like the well-being of others is their responsibility are more likely to perceive hate speech as an emergency situation and are, thus, more likely to intervene (see Latané & Darley, 1970). Moreover, due to their felt duty to protect vulnerable social groups, they are more likely to overcome any



fears of conflict and harassment that they might experience when engaging in OCI. Jost et al. (2020) argued that users conduct a cost-benefit analysis before engaging in OCI. While the possible negative consequences may be too large for many users, the commitment to solidarity may outweigh such concerns for users who strongly support solidarity citizenship norms. Overall, these expectations presuppose that, when individuals enter the online sphere, they continue to perceive themselves as citizens who are subject to the same duties and responsibilities as in the offline sphere.

Based on the aforementioned elaborations, we posited the following hypotheses:

H4–6: Strong support for solidarity citizenship norms will strengthen the effect of exposure to hate comments on the willingness to engage in disliking (H4), flagging (H5), and counterspeech (H6).

## Method

### Design

To test our hypotheses, we designed a single-factor (hate comment) between-subjects experiment embedded in an online survey. In order to generalize our findings, we used two distinct issues exemplifying verbal attacks against two distinct social groups (Issue 1: working women; Issue 2: social welfare recipients), with the participants ( $N = 337$ ) being first, randomly assigned to either of the two issues, and second, randomly assigned to either the disparaging or hateful user comments. The experimental design lacks a conventional type of control group, as both experimental groups were exposed to a condition. More specifically, one experimental group was exposed to user comments that disparaged a social group but did not use hateful language (Issue 1:  $n = 85$ ; Issue 2:  $n = 86$ ), and the other experimental group was exposed to comments that verbally attacked the same social group by using hateful language (Issue 1:  $n = 83$ ; Issue 2:  $n = 83$ ). No significant divergences were found regarding the participants' gender, age, or formal education for any of the randomized groups.

The procedure was identical for both issues and both conditions. After answering questions about their support for solidarity citizenship norms, the participants were exposed to an online news article

and the corresponding user comments. They were asked to carefully read the page and were not allowed to click the “continue” button before 30 seconds had elapsed.

### Participants

Our data was collected by a commercial online access panel (*Respondi*) in Germany. The sample included a total of 337 participants (Issue 1:  $n = 171$ , 48% female,  $M_{\text{age}} = 39.80$ ,  $SD = 11.10$ ; Issue 2:  $n = 166$ , 48.2% female,  $M_{\text{age}} = 39.23$ ,  $SD = 11.86$ ). Regarding the variance in age, we applied a quota that ensured equal representation among the age groups: 18–29, 30–39, 40–49, and 50–59. Participants above the age of 59 were excluded, as they usually have less experience with online discussion environments. Moreover, individuals who responded that they had never read user comments were screened out due to concerns of external validity. These individuals might not be familiar with comments sections and, therefore, might not be knowledgeable about functions such as flagging. Other than that, we had no specific requirements to *Respondi* with regard to participant recruitment.

### Stimuli

We created an online news article designed to resemble a typical online news site (see Appendix I). In order to prevent confounders, we blurred the name of the media outlet, the picture attached to the news article, and the profile picture and names of the commenters in the comments section. We selected two distinct issues to test whether our assumptions would hold regardless of the social group being attacked in the user comments. The news article was short and did not take a stance on the respective issues.

The first issue was about the introduction of a female quota in the boardrooms of German companies, while the second was about an increase in social welfare payments. Under each news article, we presented three user comments, the first and second of which were short and trivial (“I think they have a female quota for boardrooms/higher social welfare in other countries”; “This topic has been in the media for ages. I am tired of this debate. You read about this all the time in the

news”). The news article and the first two user comments were the same for both conditions. However, the third user comment varied (see Appendix II). In the experimental group that was exposed to user comments with disparagement, the third user comment argued against either the female quota (for Issue 1) or the increase in social welfare (for Issue 2), advancing controversial views that disparaged the respective social group. In the case of the female quota, the user comment with disparaging speech argued that women were supposed to take care of their children and the household and not aspire to a career. In the case of the welfare increase, the user comment with disparaging speech argued that social welfare recipients should just work instead of asking for more money from taxpayers. The user comments with hate speech were based on the same arguments; however, they made use of strong hate language attacking the respective social groups. In these user comments, the social groups were the subject of vulgar expletives and were threatened with violence.

## Measures

### Experimental conditions

A dummy variable was created for the experimental condition, indicating whether the participants were exposed to the user comment containing disparaging remarks (disparaging speech = 1) or the user comment characterized by hate speech (hate speech = 2).

### Moderator

To measure citizenship norms among the participants, we adapted items from the International Civic and Citizenship Education Study (ICCS). The ICCS provides a comprehensive set of citizenship norms, including items used by Dalton (2008) and Bolzendahl and Coffé (2013). Based on a confirmatory factor analysis, the ICCS citizenship norm items were divided into three dimensions, which reflected conventional citizenship, social-movement-related citizenship, and the importance of personal responsibility for citizenship (Schulz & Friedman, 2016). We selected seven items that reflected a sense of solidarity with other individuals (see Table 1), including all items in the social-

**Table 1.** Principal factor analysis of seven citizenship norms items (ICCS) reflecting solidarity.

Item	Factor loading
Participating in peaceful protests against laws believed to be unjust	.597
Participating in activities to benefit people in the city/community	.790
Taking part in activities promoting human rights	.738
Taking part in activities to protect the environment	.747
Engaging in activities to help people in less developed countries	.773
Supporting people who are worse off than you	.826
Making personal efforts to protect natural resources (e.g., through saving water or recycling waste)	.667
Eigenvalue	3.807
% of Total Variance	54.385

*N* = 337, oblimin rotation

movement-related citizenship dimension and some items in the personal responsibility dimension (Schulz & Friedman, 2016). As such, we refer to our scale as solidarity citizenship norms. The items all started with the phrase “How important are the following behaviors for being a good citizen?” and were measured on a seven-point Likert scale (1 = not important at all; 7 = very important). We conducted a principle axis factor analysis with oblimin rotation for the seven items to test for unidimensionality (Kaiser-Meyer-Olkin: 0.86). Only one factor was extracted with relatively high factor loadings (see Table 1), so we scored all seven items into a mean index ( $M = 4.95$ ,  $SD = 1.07$ ), with good internal reliability ( $\alpha = .85$ ).

### Dependent variables

All OCI-related items were measured on a seven-point Likert scale, with the participants indicating how likely they were to engage in the respective behavior (1 = very unlikely; 7 = very likely). The three variables comprised different types of OCI: (1) whether someone would report the user comment to the platform if there was an option to flag (Issue 1:  $M = 3.63$ ,  $SD = 2.22$ ; Issue 2:  $M = 2.88$ ,  $SD = 2.03$ ) and (2) whether someone would dislike the user comment (Issue 1:  $M = 4.42$ ,  $SD = 2.27$ ; Issue 2:  $M = 3.57$ ,  $SD = 2.31$ ). Moreover, an index was developed for high-threshold OCI by (3) combining four items, which asked whether the participant would be likely to write a user comment to correct the discourse (“I would call upon other users to report the comment to the platform operator,” “I would write a user comment in which

I would point out to the author of the comment to remain objective,” “I would write a user comment asking the commenter to treat other humans with respect,” “I would tell the other users to just ignore the user comment”; Issue 1: Cronbach’s  $\alpha = .88$ ,  $M = 2.99$ ,  $SD = 1.76$ ; Issue 2: Cronbach’s  $\alpha = .89$ ,  $M = 2.60$ ,  $SD = 1.68$ ). This index of high-threshold OCI is, hereafter, referred to as counterspeech.

### Data analysis

To test for effects on our three dependent variables, we ran a hierarchical ordinary least-square (OLS) regression analysis with interactions in the R-software (Version 4.0.0). The beta coefficients are unstandardized.

## Results

### Manipulation check

To check whether the manipulation of hate speech was successful, we combined six items into an index, which Naab, Kalch, and Meitz (2016) developed to test the manipulation of offensiveness in online comments (e.g., “the user comment is offensive,” “the user comment violates human rights”). We extended the index by one item – asking whether the participants perceived that the user comments presented violence as a legitimate act – and achieved good internal validity (Issue 1: Cronbach’s  $\alpha = .83$ ; Issue 2: Cronbach’s  $\alpha = .89$ ). At the end of the survey, the participants were asked to indicate whether they agreed with the various statements on a seven-point Likert scale (1 = do not agree at all; 7 = completely agree). An independent samples t-test between the two experimental groups showed that our manipulation was successful: For both issues, the hate comment was perceived as significantly more offensive than the comment without hate speech (Issue 1: hate speech condition ( $n = 86$ ):  $M = 5.36$ ,  $SD = 1.97$ ; disparaging speech condition ( $n = 85$ ):  $M = 4.00$ ,  $SD = 1.27$ ,  $t(169) = -7.17$ ,  $p < .001$ ; Issue 2: hate speech condition ( $n = 83$ ):  $M = 4.97$ ,  $SD = 1.24$ ; disparaging speech condition ( $n = 83$ ):  $M = 3.25$ ,  $SD = 1.29$ ,  $t(164) = -8.763$ ,  $p < .001$ ).

### Hypotheses tests

In H1, we expected that the participants would be more willing to dislike a user comment containing hate speech than a user comment characterized by disparaging speech. To test this hypothesis, in Step 1, we conducted a linear regression analysis for both issues. For Issue 1, the results show that the participants were not more willing to dislike a user comment attacking working women with hateful language than a user comment attacking working women with disparaging language,  $b = 0.60$ ,  $p = .082$ . Similarly, for Issue 2, there was no significant effect,  $b = .51$ ,  $p = .158$ . Thus, H1 was not supported (see Table 2, DV1, Step 1).

In H2, it was assumed that the participants would be more willing to flag a user comment containing hate speech than a user comment characterized by disparaging speech. We found that for both Issue 1,  $b = 1.89$ ,  $p < .001$ , and Issue 2,  $b = 0.89$ ,  $p = .001$ , the participants were more likely to flag hate comments than disparaging comments. Thus, H2 was supported by our data (see Table 2, DV2, Step 1). Moreover, we expected that individuals would be more likely to indicate their willingness to engage in counterspeech when exposed to user comments containing hate speech than when exposed to user comments containing disparagement (H3). This assumption was supported by our data for Issue 1,  $b = 0.73$ ,  $p = .007$ , but not for Issue 2,  $b = 0.24$ ,  $p = .369$  (see Table 2, DV3, Step 1).

As the findings for the two issues differed, for exploratory reasons, we illustrated the mean values of the different types of OCI for disparaging and hateful user comments separately for each issue. Figure 1 shows that participants were generally more willing to engage in OCI when women were attacked in user comments than when social welfare recipients were attacked. Overall, this was the case for both disparaging and hateful user comments.

In H4, we assumed that the stronger the participants’ support for solidarity citizenship norms, the more willing they would be to dislike user comments containing hate speech compared to user comments characterized by disparaging speech. To test this assumption, in Step 2, we added an interaction term that included the user comment and solidarity citizenship norms to the



**Table 2.** Effects of hate comments and solidarity citizenship norms on disliking, flagging and counterspeech (Hierarchical regression analysis).

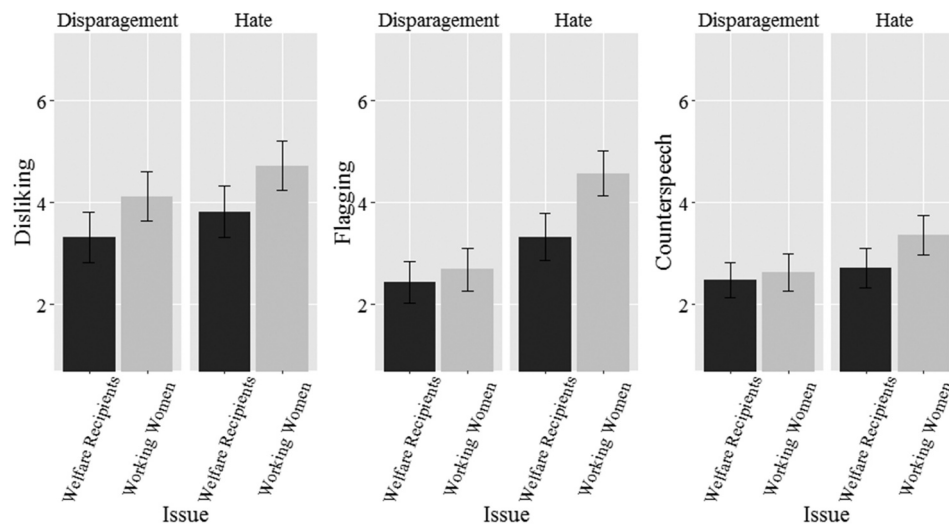
	Issue 1: Working Women				Issue 2: Welfare Recipients			
	$R^2$	$b$	$SE$	$p$	$R^2$	$b$	$SE$	$p$
<b>DV1: Disliking</b>								
<b>Step 1</b>	.07				.14			
Constant		1.01	1.00	.312		-1.17	0.96	.222
Hate comment <sup>a</sup>		0.60	0.35	.082		0.51	0.36	.158
<b>Step 2</b>	.07				.14			
Constant		1.98	2.76	.474		-0.83	2.48	.738
Solidarity citizenship norms		0.28	0.53	.600		0.72	0.49	.148
Hate comment <sup>a</sup>		0.08	1.66	.961		0.37	1.58	.814
Solidarity citizenship norms * hate comment <sup>a</sup>		0.12	0.32	.706		0.05	0.31	.881
<b>DV2: Flagging</b>								
<b>Step 1</b>	.22				.14			
Constant		-1.25	.90	.163		-1.41	0.84	.096
Hate comment <sup>a</sup>		1.89***	0.31	.000		0.89**	0.31	.004
<b>Step 2</b>	.28				.15			
Constant		6.98**	2.38	.004		2.08	2.16	.338
Solidarity citizenship norms		-1.24**	0.46	.008		-0.13	0.43	.765
Hate comment <sup>a</sup>		-3.23**	1.43	.025		-1.39	1.38	.314
Solidarity citizenship norms * hate comment <sup>a</sup>		1.03***	0.28	.000		0.48	0.28	.082
<b>DV3: Counterspeech</b>								
<b>Step 1</b>	.11				.11			
Constant		-0.33	0.76	.66		-0.35	0.71	.623
Hate comment <sup>a</sup>		0.73**	0.26	.007		0.24	0.26	.369
<b>Step 2</b>	.13				.11			
Constant		3.49	2.07	.093		0.19	1.84	.920
Solidarity citizenship norms		-0.33	0.40	.404		.40	.37	.271
Hate comment <sup>a</sup>		-1.60	1.24	.198		-0.06	-1.17	.957
Solidarity citizenship norms * hate comment <sup>a</sup>		0.48*	0.24	.049		0.07	0.23	.753

Issue 1:  $n = 171$ , Issue 2:  $n = 166$ , \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .<sup>a</sup> Disparaging Speech = 1, Hate Speech = 2

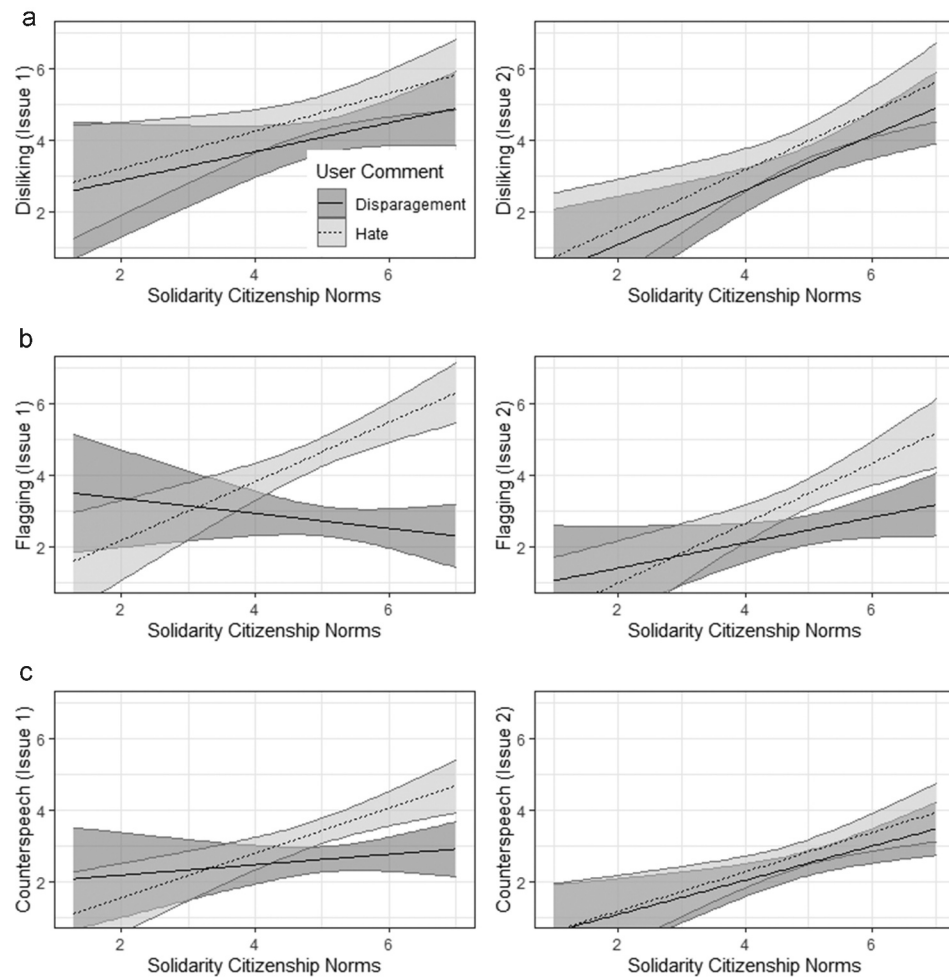
model. The interaction term was insignificant for both Issue 1,  $b = 0.12$ ,  $p = .706$ , and Issue 2,  $b = 0.05$ ,  $p = .881$ . Consequently, H4 was not supported by our data (see Table 2, DV1, Step 2). In H5, we expected that the stronger the individuals' support for solidarity citizenship norms, the more willing they would be to flag a hate comment compared to a disparaging comment. While we found a significant interaction term for Issue 1,  $b = 1.03$ ,  $p < .001$ , for Issue 2, the interaction term was not significant,  $b = 0.48$ ,  $p = .082$  (see Table 2, DV2, Step 2). In H6, we expected that the likelihood of engaging in counterspeech would increase with stronger citizenship norms. While we found a significant interaction term for Issue 1,  $b = 0.48$ ,  $p = .049$ , this was not the case for Issue 2,  $b = 0.07$ ,  $p = .753$  (see Table 2, DV3, Step 2).

For a more nuanced understanding of these findings, we visualized the interaction (see Figure 2). The illustration shows that individuals with strong support for solidarity citizenship norms were generally more willing to dislike disparaging and hateful user comments than individuals with weak support for solidarity citizenship norms (see

Figure 2, Section a). However, neither participants with strong support for solidarity citizenship norms nor participants with weak support for solidarity citizenship norms distinguished between disparaging and hateful user comments. In contrast, Section b in Figure 2 illustrates that for both Issues, the willingness to flag a hate comment (compared to a disparaging comment) increased more strongly for participants with stronger support for solidarity citizenship norms than for participants with weaker support for solidarity citizenship norms. The interaction term was, however, only significant for Issue 1. Lastly, Figure 2 shows that for Issue 1, the willingness to engage in counterspeech against hate comments (compared to disparaging comments) increased more strongly for participants with strong support for solidarity citizenship norms than for participants with weak support for solidarity citizenship norms (see Figure 2, Section c). For Issue 2, individuals with stronger support for solidarity citizenship norms were generally more likely to engage in counterspeech, although they did not distinguish between hateful and disparaging user comments.



**Figure 1.** Mean values for willingness to engage in disliking (a), flagging (b), and counterspeech (c).  $N = 337$



**Figure 2.** Interactions between hate comments and solidarity citizenship norms on disliking (a), flagging (b), and counterspeech (c).  $N = 337$

## Discussion

As hate speech has become a pervasive phenomenon in communicative online spaces, media platforms have increasingly relied on users to take action (Gillespie, 2018). Consequently, scholars have started to examine how contextual, individual, and content characteristics may encourage or discourage such user engagement (e.g., Kalch & Naab, 2017; Leonhard et al., 2018; Watson et al., 2019; Wilhelm et al., 2019). One aim of the present study was to find further empirical support for the assumption that OCI is more likely when comments contain strong, hateful language than when they contain subtle, disparaging language (e.g., Kalch & Naab, 2017; Leonhard et al., 2018). The other aim was to advance the knowledge about potential individual predictors by testing whether support for solidarity citizenship norms might moderate the effect of exposure to hate comments on the willingness to engage in OCI.

## Findings

As predicted by our hypotheses, the willingness to engage in flagging and counterspeech increased when participants were exposed to comments characterized by hate compared to comments that disparaged a social group without the use of abusive language. These findings provide further support for previous studies (Jost et al., 2020; Kalch & Naab, 2017; Leonhard et al., 2018; Wilhelm et al., 2019). However, we found a significant effect when user comments attacked working women but not when they attacked social welfare recipients. Thus, these findings indicate that the willingness to engage in OCI against hate comments depends on the attacked social group. One explanation might be that social welfare recipients are perceived as a less discriminated social group than women. As there has been a global public debate about women being the victims of hate speech in the digital sphere, users may be particularly sensitive to this issue and find hate comments against women particularly alarming. In contrast, hate speech against social welfare recipients has not been an issue that has received much public attention, and the study participants were, thus, unlikely to perceive it as a major societal problem. Overall, with regard to both disparaging and hateful comments, participants were more

willing to engage in OCI in the case of women than in the case of social welfare recipients. There is a need for more research on how users perceive both hate speech and disparaging speech toward different social groups and how this may, in turn, influence OCI. Another finding of the present study was that, for Issue 1 and 2, the participants were not more willing to dislike hateful comments than disparaging comments. This finding supports Kalch and Naab's (2017) results, indicating that users do not regard disliking as an appropriate measure against hate comments.

By investigating the moderating effect of citizenship norms on OCI, we provide new empirical insight into research that deals with the relationship between digital media use, citizenship norms, and participation (Feezell et al., 2016; Ohme, 2018; Thorson, 2015). We argued that support for citizenship norms that emphasize solidarity with others may moderate the effect of exposure to hate comments on the willingness to engage in OCI. After all, individuals who engage in OCI against hate speech come to the defense of vulnerable social groups and provide them with social support. In sum, we found that the stronger the participant's support for solidarity citizenship norms, the higher the willingness to engage in flagging or counterspeech when working women were attacked with hate speech (compared to disparaging speech). However, solidarity citizenship norms did not have a similar moderating effect with regard to social welfare recipients. Again, individuals who supported solidarity citizenship norms probably perceived women as the more discriminated social group. The felt responsibility of showing solidarity may generally be stronger if a social group is perceived as particularly vulnerable in the digital public sphere, which is why participants with strong support for solidarity citizenship norms felt more obliged to intervene in the case of women than in the case of social welfare recipients.

While there was no significant interaction term for Issue 2, Figure 2 indicates that the stronger the support for solidarity citizenship norms, the higher the willingness to engage in flagging or counterspeech with regard to both disparaging and hateful comments. Thus, participants with strong support for solidarity citizenship norms seemed to feel generally more obliged to intervene than participants with weak support for solidarity citizenship norms. Similarly, for Issues 1 and 2, the stronger the support for solidarity

citizenship norms, the more likely the participants were to dislike both hateful and disparaging user comments. This finding indicates that individuals with strong support for solidarity citizenship norms may perceive the dislike button as a convenient tool to show solidarity with attacked social groups, even if these attacks are more subtle. In other words, the dislike button does not seem to be perceived as a tool specifically for fighting hate speech.

In sum, this study contributes to the small but growing body of research on the forms and factors of OCI. Solidarity citizenship norms were shown to be a significant catalyst of OCI, indicating that individuals with strong support for solidarity citizenship norms do not abandon their perceived responsibilities as citizens when they enter the digital space. Thus, communicative online spaces are not perceived as detached from the non-digital spheres of life; rather, they are perceived as spaces where similar norms apply.

Lastly, we tested the effects of exposure to hate speech and solidarity citizenship norms on the willingness to engage in OCI in the communicative online space of comment sections beneath news articles. Hence, the transferability of the results to social media environments is limited. In social media environments, users have a personal profile, commonly use their real name, and have a network with friends and acquaintances. The degree of felt anonymity, thus, tends to be lower in social media environments than in comment sections. Consequently, in social media environments, users may be more reluctant to push social buttons or engage in counterspeech because they may be judged or become entangled in conflicts with people they know. However, individuals may be encouraged to engage in these public activities if they expect social recognition from their peers. Further research could explore how the degree of perceived anonymity on different media platforms affects individuals' willingness to engage in OCI and how this may interact with individual factors, such as support for citizenship norms.

## Limitations

This study has some limitations, which can be addressed in future studies. First, the measurement of our dependent variable was based on self-reports. Thus, we only measured whether individuals thought

they would likely engage in OCI instead of whether they actually did so. In the context of this study, social desirability may have had a considerable effect on the participants' answers. The effect of social desirability may have even interacted with the effect of solidarity citizenship norms on OCI, as individuals who believe that good citizens should stand up for others may also think that they should intervene in the case of hate comments. Thus, individuals with high levels of support for solidarity citizenship norms may perceive engagement in OCI as especially socially desirable. For a more realistic environment, future studies should develop mock sites on which participants could flag, dislike, or respond to user comments (similar to Kalch & Naab, 2017; Naab et al., 2016) or, even better, conduct a field experiment (Munger, 2017). Endeavors to increase the ecological validity of the experimental setting through field experiments would increase the robustness of our findings.

Second, we did not run a pretest on the perceived valence of the news article that was shown to the participants or ask about such perceptions in the web survey. Thus, even though the article was very short, and we aimed to make it appear un-slanted, we cannot be completely certain that we were successful. Yet, even if the participants had perceived a slight slant, we do not believe that this would have affected the study outcomes. After all, it is unlikely that participants would have perceived hate speech against a social group as more or less legitimate just because it had a slightly positive or negative view of the female's quota or social welfare. However, to avoid this possibility, pretests would have been meaningful.

Third, scholars have argued that, as citizenship norms are commonly measured as support for certain types of political participation, predicting political participation through citizenship norms is problematic because the concepts are closely related (Ohme, 2018). However, this concern has only minor relevance in the context of this study as we did not measure the effect of participants' support for a certain political behavior on participants' engagement in that behavior. Instead, OCI involves a very different set of activities than the activities addressed by solidarity citizenship norms. Accordingly, the correlations between the different types of OCI and solidarity citizenship norms were significant but not very strong (below .3, see Appendix III). However, we cannot

completely exclude the possibility of a spurious relationship. Since support for solidarity citizenship norms may also reflect whether the participants engage in such activities, the relationship between solidarity citizenship norms and OCI may have been caused by political and social engagement as a third variable. However, scales that measure citizenship norms without referring to specific activities are generally rare (Ohme, 2018) and there is no such scale for solidarity citizenship norms specifically. Thus, to solve the problem of potential spurious effects, more conceptual work on how to measure citizenship norms as a predictor for different types of political participation is necessary. Nevertheless, some empirical studies have found no relationship between citizenship norms and political participation (Bolzendahl & Coffé, 2013; Copeland & Feezell, 2017), indicating that the measurement of citizenship norms is distinct from the measurement of political participation. In other words, individuals' expectations about a good citizen's ideal actions do not necessarily imply that those individuals live up to their own expectations.

Fourth, we compared the effect of a blunt hate comment to that of a disparaging non-hate comment. While, in our study, the degrees of deviance between these two comments were strong, in reality, many hate comments might be more subtle, and the line between hate comments and those characterized somewhat innocuously as simply disparaging might be less obvious. Thus, future studies should investigate the effect of the nuances of hate speech and analyze whether individuals who believe in solidarity citizenship norms are also more likely to intervene in cases of less blatant hate speech.

### Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Funding

This work was supported by the Bundesministerium für Bildung und Forschung [16DII14].

### Notes on contributors

**Marlene Kunst** is a research associate and PhD candidate at the Freie Universität Berlin/Weizenbaum Institute for the Networked Society, Berlin. Her research focuses on political communication, media effects in polarized digital media environments, and political participation online.

**Pablo Porten-Cheé** (PhD, University of Düsseldorf) is head of the research group "Digital Citizenship" at the Weizenbaum Institute for the Networked Society, Berlin and postdoctoral fellow at the Freie Universität Berlin, Germany. His research spans political communication, political participation, and effects of political content under online conditions.

**Martin Emmer** is professor for media and communication studies at Freie Universität Berlin and founding director (2017-19) of the Weizenbaum Institute for the Networked Society, Berlin. His research interests include political online communication, digital citizenship, and digital methods.

**Christiane Eilders**, is a professor of communication studies at the Heinrich-Heine-University Düsseldorf, Germany. Her main research interests are in the field of political communication, in particular public sphere and civil society, public opinion formation, participation and deliberation online.

### ORCID

Marlene Kunst  <http://orcid.org/0000-0003-0729-9749>

Pablo Porten-Cheé  <http://orcid.org/0000-0002-7594-9363>

Martin Emmer  <http://orcid.org/0000-0002-0722-132X>

### References

- Bennett, L. W. (2008). Changing citizenship in the digital age. In L. W. Bennett (Ed.), *Civic life online: Learning how digital media can engage youth* (pp. 124). Cambridge, MA: MIT Press.
- Boeckmann, R. J., & Liew, J. (2002). Hate speech: Asian American students' justice judgments and psychological responses. *Journal of Social Issues*, 58(2), 363–381. doi:10.1111/1540-4560.00265
- Bolzendahl, C., & Coffé, H. (2013). Are 'good' citizens 'good' participants? Testing citizenship norms and political participation across 25 nations. *Political Studies*, 61(1), 45–65. doi:10.1111/1467-9248.12010
- Chang, W.-C. (2016). Culture, citizenship norms, and political participation: Empirical evidence from Taiwan. *Japanese Journal of Political Science*, 17(2), 256–277. doi:10.1017/S1468109916000062
- Chen, G. M. (2017). *Nasty talk: Online incivility and public debate*. Cham: Palgrave Macmillan.
- Copeland, L. (2014). Conceptualizing political consumerism: How citizenship norms differentiate boycotting from buycotting. *Political Studies*, 62(1), 172–186. doi:10.1111/1467-9248.12067



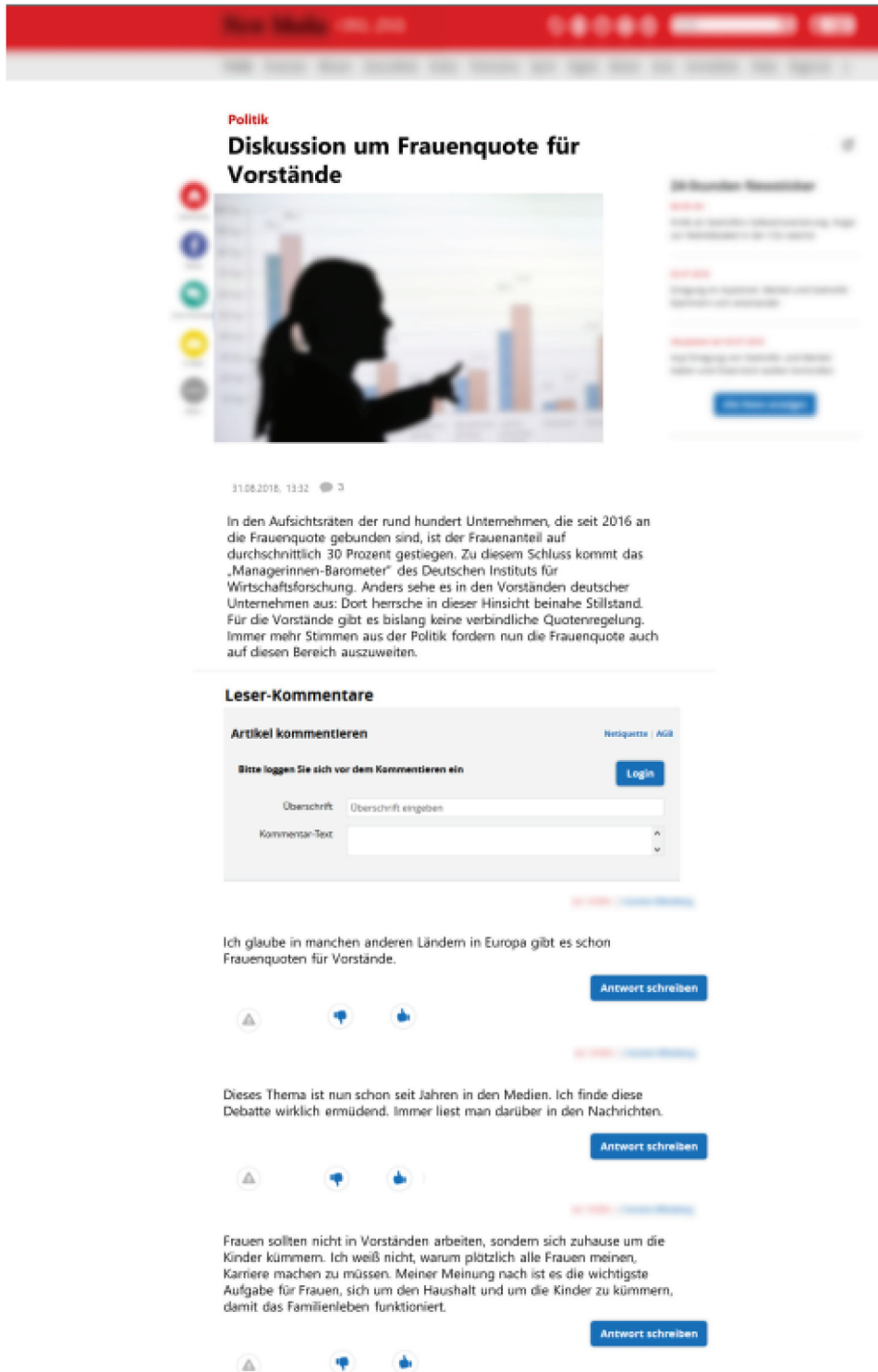
- Copeland, L., & Feezell, J. T. (2017). The influence of citizenship norms and media use on different modes of political participation in the US. *Political Studies*, 65(4), 805–823. doi:10.1177/0032321717720374
- Dalton, R. J. (2006). Citizenship norms and political participation in America good news is . . . the bad news is wrong. *The Center for Democracy and Civil Society*(CDACS Occasional Paper 2006–01). <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.495.4400&rep=rep1&type=pdf>
- Dalton, R. J. (2008). Citizenship norms and the expansion of political participation. *Political Studies*, 56(1), 76–98. doi:10.1111/j.1467-9248.2007.00718.x
- Delgado, R., & Stefancic, J. (2019). *Understanding words that wound*. Abingdon: Routledge.
- Erjavec, K., & Kovačič, M. P. (2012). “‘You don’t understand, this is a new war!’ analysis of hate speech in news web sites’ comments”. *Mass Communication and Society*, 15(6), 899–920. doi:10.1080/15205436.2011.619679
- Feezell, J. T., Conroy, M., & Guerrero, M. (2016). Internet use and political participation: Engaging citizenship norms through online activities. *Journal of Information Technology & Politics*, 13(2), 95–107. doi:10.1080/19331681.2016.1166994
- Gardiner, B., Mansfield, M., Anderson, I., Holder, J., Louter, D., & Ulmanu, M. (2016). The dark side of guardian comments. *The Guardian Online*. <https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments>
- Gillespie, T. (2018). *Custodians of the Internet. platforms, content moderation, and the hidden decisions that shape social media*. New Haven & London: Yale University Press.
- Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-Muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology*, 11, 143–160. <http://dx.doi.org/10.5281/zenodo.495778>
- Hooghe, M., Oser, J., & Marien, S. (2016). A comparative analysis of ‘good citizenship’: A latent class analysis of adolescents’ citizenship norms in 38 countries. *International Political Science Review*, 37(1), 115–129. doi:10.1177/0192512114541562
- Jost, P., Ziegele, M., & Naab, T. K. (2020). Klicken oder tippen? Eine analyse verschiedener interventionsstrategien in unzi- vilen online-diskussionen auf facebook. *Zeitschrift Für Politikwissenschaft*, 105(1), 231. doi:10.1007/s41358-020-00212-9
- Kalch, A., & Naab, T. (2017). Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *Studies in Communication | Media*, 6(4), 395–419. doi:10.5771/2192-4007-2017-4-395
- Kim, M. (2018). How does facebook news use lead to actions in South Korea? The role of Facebook discussion network heterogeneity, political interest, and conflict avoidance in predicting political participation. *Telematics and Informatics*, 35(5), 1373–1381. doi:10.1016/j.tele.2018.03.007
- Knobloch-Westerwick, S., Sharma, N., Hansen, D. L., & Alter, S. (2005). Impact of popularity indications on readers’ selective exposure to online news. *Journal of Broadcasting & Electronic Media*, 49(3), 296–313. doi:10.1207/s15506878jobem4903\_3
- Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn’t he help?* New York: Appleton-Century-Crofts.
- 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 counter-argue against hate speech on Facebook. *Studies in Communication | Media*7(4), 555–579. doi:10.5771/2192-4007-2018-4-555.
- Ley, H. (2018). #ICHBINHIER. Zusammen gegen Fake News und Hass im Netz. [IAMHERE. United against Fake News and Hate Online]. Köln: DuMont Buchverlag GmbH.
- Munger, K. (2017). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*39(3), 629–649. doi:10.1007/s11109-016-9373-5
- Naab, T., Kalch, A., & Meitz, T. (2016). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, 20(2), 777–795. doi:10.1177/1461444816670923
- Naab, T. K. (2016). Der Sanktionsbedarf von Facebook-Inhalten aus Sicht von Nutzerinnen und seine Determinanten. *Medien & Kommunikationswissenschaft*, 64(1), 56–73. doi:10.5771/1615-634X-2016-1-56
- Ohme, J. (2018). Updating citizenship? The effects of digital media use on citizenship understanding and political participation. *Information, Communication & Society*, 1–26. doi:10.1080/1369118X.2018.1469657
- Porten-Cheé, P., Kunst, M., & Emmer, M. (2020). Online Civic Intervention: A New Form of Political Participation Under Conditions of a Disruptive Online Discourse. *International Journal of Communication*14, 514–534.
- Schulz, W., & Friedman, T. (2016). Scaling procedures for ICCS questionnaire items. In W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.) *ICCS 2016. Technical Report*. The International Association for the Evaluation of Educational Achievement (IEA). <https://www.iea.nl/publications/technical-reports/iccs-2016-technical-report>
- Singer, J. B. (2014). User-generated visibility: Secondary gate-keeping in a shared media space. *New Media & Society*, 16(1), 55–73. doi:10.1177/1461444813477833
- Sirsch, J. (2013). Die Regulierung von Hassrede in liberalen Demokratien [the regulation of hate speech in liberal democracies]. In J. Meibauer (Ed.), *Linguistische Untersuchungen. Hassrede - von der Sprache zur Politik*. Giessen: Gießener Elektronische Bibliothek.
- Theiss-Morse, E. (1993). Conceptualizations of good citizen-ship and political participation. *Political Behavior*, 15(4), 355–380. doi:10.1007/BF00992103
- Theocharis, Y., & van Deth, J. W. (2017). *Political participation in a changing world: Conceptual and empirical*

- challenges in the study of citizen engagement*. Abingdon: Routledge.
- Thorson, K. (2015). Sampling from the civic buffet: Youth, new media and do-it-yourself citizenship. In H. Gilde Zúñiga (Ed.), *New technologies & civic engagement: New agendas in communication* (pp. 3–23). Abingdon: Routledge.
- van Deth, J. W. (2007). Norms of citizenship. In R. J. Dalton & H. Klingemann (Eds.), *The Oxford handbook of political behavior* (pp. 1–19). Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199270125.003.0021
- Watson, B. R., Peng, Z., & Lewis, S. C. (2019). Who will intervene to save news comments? Deviance and social control in communities of news commenters. *New Media & Society*, 21(8), 1840–1858. doi:10.1177/1461444819828328
- Wilhelm, C., & Joeckel, S. (2018). Gendered morality and backlash effects in online discussions: An experimental study on how users respond to hate speech comments against women and sexual minorities. *Sex Roles*, 80(7–8), 381–392. doi:10.1007/s11199-018-0941-5
- Wilhelm, C., Joeckel, S., & Ziegler, I. (2019). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *communication research*, 3. doi:10.1177/0093650219855330

## Appendix A

### Screenshot of stimulus material

Illustration 1. Example of stimulus.



**Politik**

## Diskussion um Frauenquote für Vorstände

31.08.2018, 13:32

In den Aufsichtsräten der rund hundert Unternehmen, die seit 2016 an die Frauenquote gebunden sind, ist der Frauenanteil auf durchschnittlich 30 Prozent gestiegen. Zu diesem Schluss kommt das „Managerinnen-Barometer“ des Deutschen Instituts für Wirtschaftsforschung. Anders sehe es in den Vorständen deutscher Unternehmen aus: Dort herrsche in dieser Hinsicht beinahe Stillstand. Für die Vorstände gibt es bislang keine verbindliche Quotenregelung. Immer mehr Stimmen aus der Politik fordern nun die Frauenquote auch auf diesen Bereich auszuweiten.

### Leser-Kommentare

Artikel kommentieren [Netiquette](#) [AGB](#)

Bitte loggen Sie sich vor dem Kommentieren ein [Login](#)

Überschrift:

Kommentar-Text:

[Antwort schreiben](#)

Ich glaube in manchen anderen Ländern in Europa gibt es schon Frauenquoten für Vorstände.

[Antwort schreiben](#)

Dieses Thema ist nun schon seit Jahren in den Medien. Ich finde diese Debatte wirklich ermüdend. Immer liest man darüber in den Nachrichten.

[Antwort schreiben](#)

Frauen sollten nicht in Vorständen arbeiten, sondern sich zuhause um die Kinder kümmern. Ich weiß nicht, warum plötzlich alle Frauen meinen, Karriere machen zu müssen. Meiner Meinung nach ist es die wichtigste Aufgabe für Frauen, sich um den Haushalt und um die Kinder zu kümmern, damit das Familienleben funktioniert.

[Antwort schreiben](#)

## Appendix B

### Experimental stimulus: Disparaging comments and hate comments

#### Issue 1: Working women

##### Disparaging comment

Women should not work on boards of directors and should take care of their children at home. I don't know why suddenly all women think they have to have a career. In my opinion, the most important task for women is to take care of the household and the children so that family life can work.

##### Hate comment

These bitches shouldn't be on the board of directors, but should take care of the children at home. Such career bitches need to get their faces smashed so that they understand that they have to take care of the household and children. But these disgusting cunts don't care about the family!

#### Issue 2: Welfare recipients

##### Disparaging comment

In my opinion, the unemployed should not try to get more money from the state but go to work to earn their own money.

It is better if they don't get more money for doing nothing. After all, welfare is financed by hard-working taxpayers.

##### Hate comment

Welfare recipients should get their heads straight. They just sit around at home, eat fast food and drink themselves to death. Disgusting! Those who try to get more money deserve to get their faces smashed. It's better if they do not get too much money because we taxpayers pay with our hard-earned money for this miserable rabble.

## Appendix C

**Table A1.** Pearson correlation between solidarity citizenship norms and OCI measures.

Variable	1	2	3	4
1. Solidarity citizenship Norms	1.00**			
2. Disliking	0.28**	1.00**		
3. Flagging	0.20**	0.47**	1.00**	
4. Counterspeech	0.27**	0.61**	0.58**	1.00**

$N = 337$ , \* $p < .05$ . \*\* $p < .01$ .