SCM
Studies in Communication and Media

FULL PAPER

Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments

Journalistische Gegenrede in Kommentarbereichen: Strukturen, Determinanten und mögliche Konsequenzen der interaktiven Moderation von inzivilen Nutzerkommentaren

Marc Ziegele, Pablo Jost, Marike Bormann & Dominique Heinbach

Marc Ziegele (Ass. Prof. Dr.), Department of Social Sciences, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany; Contact: ziegele(at)hhu.de

Pablo Jost (M.A.), Department of Communication, Johannes Gutenberg University of Mainz, Jakob-Welder-Weg 12, 55128 Mainz, Germany; Contact: pablo.jost(at)uni-mainz.de

Marike Bormann (M.A.), Department of Social Sciences, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany; Contact: bormann(at)phil.hhu.de

Dominique Heinbach (M.A.), Department of Social Sciences, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany; Contact: heinbach(at)phil.hhu.de

FULL PAPER

Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments

Journalistische Gegenrede in Kommentarbereichen: Strukturen, Determinanten und mögliche Konsequenzen der interaktiven Moderation von inzivilen Nutzerkommentaren

Marc Ziegele, Pablo Jost, Marike Bormann & Dominique Heinbach

Abstract: Incivility in online user discussions is discussed as a significant challenge for democratic societies. Interactive journalistic moderation is seen as a promising strategy to deal with and prevent online incivility. Such moderation occurs, for example, when journalists publicly respond to uncivil comments and ask their authors to discuss more civilly. This study, based on a quantitative content analysis of 9,763 user and moderation comments on the Facebook sites of 15 German news outlets, investigated the patterns, determinants, and potential effects of interactive moderation. Results show that so-called public-level incivility (e.g., stereotypes, threats of violence) in users' initial comments was associated with more interactive journalistic moderation, and that journalists used different styles when responding to these comments. Different moderation styles of initial comments were then related to the presence of incivility in users' subsequent reply comments in opposite directions: A sociable moderation style decreased, and a regulative style increased the level of incivility in the reply comments.

Keywords: User comments, incivility, counter-voices, moderation, content analysis

Zusammenfassung: Der raue und zuweilen hasserfüllte Umgangston von Nutzerinnen und Nutzern in Online-Diskussionen stellt demokratische Gesellschaften vor eine Herausforderung. Interaktive journalistische Moderation gilt als eine vielversprechende Maßnahme, um das hohe Maß an sogenannter Online-Inzivilität einzudämmen. Im Rahmen einer solchen Moderation reagieren Journalisten oder Community-Manager öffentlich mit Gegenrede auf inzivile Kommentare und bitten die Verfasser zum Beispiel, ihren Umgangston zu mäßigen. Unsere Studie untersucht die Strukturen, Determinanten und möglichen Auswirkungen von interaktiver Moderation mittels einer quantitativen Inhaltsanalyse von 9.763 Nutzer- und Moderationskommentaren auf den Facebook-Seiten von 15 deutschen Nachrichtenmedien. Die Ergebnisse zeigen, dass Kommentare, die sogenannte public-level incivility (u. a. Stereotypen, Androhung von Gewalt) enthalten, häufiger moderiert werden und dass die Moderatoren mit unterschiedlichen Stilen auf die Kommentare antworten. Zudem zeichnet sich ab, dass ein geselliger Moderationsstil die Inzivilität der Folgekommentare reduziert, während ein regulativer Moderationsstil die Inzivilität der anschließenden Diskussion sogar noch erhöht.

Schlagwörter: Nutzerkommentare, Gegenrede, Inzivilität, Moderation, Inhaltsanalyse

1. Introduction

Public user comments on the websites and Facebook pages of established news media outlets are a popular element of digital online communication (Stroud, Scacco, Muddiman, & Curry, 2015). Many news consumers use comment sections to learn about the issue-related attitudes of other users or to voice their own opinions towards news topics and other user comments (Rowe, 2015b; Springer, Engelmann, & Pfaffinger, 2015). For German online users, reading comments is now almost as widespread as reading printed newspapers (Ziegele, Köhler, & Weber, 2017): Forty-one percent of German onliners read user comments at least once a week, and 50 percent read printed newspapers on a regular basis. Thirty percent contribute own comments at least once a month. In samples representative of the U.S. population, 24 percent of the participants answered to write user comments on news sites weekly (Newman, Fletcher, & Kalogeropoulos, Levy, & Nielsen, 2017), while 35 percent read the comments of others but did not participate themselves (Stroud, van Duyn, & Peacock, 2016).

The quality of user comments is often assessed through the lens of *deliberation theory* (e.g., Manosevitch & Walker, 2009; Rowe, 2015b; Ruiz et al., 2011). The deliberation framework sketches a public sphere that can be accessed by everyone, and in which citizens discuss social and political issues in a rational, reciprocal, and respectful manner (Gastil, 2008; Habermas 1996). From this theoretical perspective, comment sections could be a promising forum for an open, non-discriminatory, and constructive exchange of opinions between citizens on socially significant issues (Friess & Eilders, 2015; Ruiz et al., 2011). However, researchers, journalists, and politicians have raised concerns regarding the deliberative quality of user comments. Many comments are not constructive, respectful, and result-oriented. Rather, they include a high degree of *incivility* (Coe, Kenski, & Rains, 2014). Content analyses have demonstrated that between 20 and 40 percent of user comments on various U.S. and German news sites include some degree of incivility (Coe et al., 2014; Santana, 2014; Ziegele, Quiring, Esau, & Friess, 2018).

Uncivil comments are problematic because they can undermine democratic values and lead to attitude polarization (Anderson, Brossard, Scheufele, Xenos, & Ladwig, 2014). Moreover, they can increase aggressive cognitions and stereotypical attitudes among their readers and have a negative impact on the perceived news quality of established news media (Hsueh, Yogeeswaran, & Malinen, 2015; Prochazka, Weber, & Schweiger, 2018). In the long run, overly uncivil discussions prevent users from writing comments, make journalists feel worries and anger, and can make news outlets shut down their comment sections altogether (Obermaier, Hofbauer, & Reinemann, this issue; Springer et al., 2015; Stroud et al., 2015; Stroud et al., 2016).

Incivility in online user discussions is therefore a significant challenge for democratic societies and developing strategies to deal with incivility is considered as an important task for communication research. One promising approach to fostering the development of civil norms and behaviors and to improving the discus-

sion atmosphere in comment sections without overly limiting free speech is *interactive journalistic moderation* (Meyer & Carey, 2014; Stroud et al., 2015; Ziegele, & Jost, 2016). Interactive journalistic moderation occurs when journalists (or community managers, respectively) publicly respond to the comments of their users. Such moderation, in the case of uncivil comments, can be conceptualized as a manifestation of journalistic counter-voices in comment sections.

The current study investigates the patterns and the potential effects of interactive journalistic moderation of uncivil user comments across 15 German news sites on Facebook. More specifically, it examines whether uncivil comments are related to increased interactive journalistic moderation and how journalists respond to different levels of incivility. The analysis reveals different response styles journalists use when interactively engaging with the uncivil comments of their readers. Furthermore, we investigate how the presence and styles of interactive moderation of initial comments relate to the level of incivility in the following reply comments. From the findings, we draw conclusions regarding the effectiveness of journalistic counter-voices and their potential to increase the civility of online discussions.

2. Incivility in user comments

As the decision of what is civil and uncivil is subjectively shaped (Herbst, 2010), incivility is a "notoriously difficult term to define" (Coe et al., 2014, p. 660). Therefore, and despite increasing academic attention towards the phenomenon, researchers have mentioned the lack of an agreed-upon definition as well as a unifying model of incivility (Muddiman, 2017; Stryker, Conway, & Danielson, 2016). Hence, achieving consensus about where to draw the line between civil and uncivil discourse is a complex problem. Recent efforts have been based on the distinction between true incivility, which undermines the ideal of deliberative discussions, and mere negativity as an inevitable characteristic of disagreement (e.g., Massaro, & Stryker, 2012). More specifically, mere negativity and disagreement qualify as fundamental parts of political discourse because they function as indicators for diverse viewpoints (Stromer-Galley, 2007). Negativity, however, only remains functional for democracies if it is presented in a respectful and polite manner (Herbst, 2010). Thus, negativity alone does not constitute incivility, but negativity combined with a dismissive, disrespectful, aggressive, and hostile tone makes a statement uncivil (Coe et al., 2014; Hwang, Kim, & Kim, 2016). From this perspective, incivility and uncivil behavior can be defined as the "expression of disagreement by denying and disrespecting […] the opposing views" (Hwang et al., 2016, p. 5).

In user comments, incivility encompasses rhetorical and stylistic elements such as insulting vocabulary, ad hominem attacks, or verbal intimidation on the one hand (Coe et al., 2014; Ziegele, & Jost, 2016). These forms can violate social norms in communication processes, such as interpersonal politeness norms (e.g., Mutz, 2015). On the other hand, incivility in online comments can also appear as a "set of behaviors that threaten democracy, deny people their personal freedoms, and stereotype social groups" (Papacharissi, 2004, p. 267). Examples of such incivility in-

clude racism, sexism, attacking people for belonging to certain social or ethnic groups, or threatening other individuals´ rights (Kalch & Naab, 2017; Papacharissi, 2004).[1] These uncivil behaviors do not only violate interpersonal politeness norms, but rather norms of "collective politeness" (Papacharissi, 2004, p. 267).

Therefore, according to Papacharissi (2004), a basic distinction can be drawn between incivility and *impoliteness*, with "politeness as etiquette-related, and civility as respect for the collective traditions of democracy" (p. 260). She argues that politeness is a necessary but not sufficient condition for civility and that civility cannot be confined to impoliteness since "robust and heated discussions" (p. 260) can also be advantageous for democratic discourse. Studies that distinguish between impoliteness and incivility typically reveal that impoliteness is relatively widespread in online discussions, whereas incivility occurs less frequently (Papacharissi, 2004; Rowe, 2015a).

Based on the distinction between the violation of interpersonal and collective norms, Muddiman (2017) conceptualized a two-dimensional model of incivility, in which personal-level incivility, on the one hand, encompasses different forms of impoliteness. Public-level incivility, on the other hand, describes those forms of incivility that relate to "violating norms of political and deliberative processes." (Muddiman, 2017, p. 3183). By demonstrating that individuals perceive both kinds of norm violations as uncivil, Muddiman (2017) validated incivility as a two-dimensional construct.

Regarding the types of messages that fall into the two categories, Papacharissi (2004) considered comments as impolite when they contain name-calling, vulgarity, or less obvious types of impoliteness such as sarcasm or using all-caps to reflect shouting, for example. Muddiman's (2017) types of personal-level incivility partly overlap with this operationalization and include insulting language and name-calling, obscene language, as well as emotional language and displays, such as anger or yelling. In contrast, Papacharissi (2004) classified a message as uncivil when it threatens democracy or other individuals' rights, or when it assigns stereotypes. Muddiman's (2017) conceptualization of public-level incivility encompasses messages including misinformation and accusations of lying, ideological extremity and lack of comity, lack of compromise, as well as nonpublic acts.

Based on these considerations, and following Muddiman's (2017) wording, the current study conceptualizes incivility as a two-level model, drawing a distinction between a) types of incivility that violate norms of interpersonal politeness and b) types of incivility that violate norms of public political and deliberative processes. Personal-level incivility includes comments that use insults (insulting language and name-calling), profanity (vulgarity and obscenity), screaming (shouting and yelling; Muddiman, 2017; Papacharissi, 2004), or sarcasm or cynicism (Papacharissi, 2004). Public-level incivility encompasses comments that assign antagonistic stereotypes, use threats of violence (Papacharissi, 2004), or accuse others of lying (Muddiman, 2017).

1 It needs to be mentioned, however, that studies vary significantly regarding the number and types of uncivil behaviors included.

This operationalization, although not mirroring the full spectrum of impolite and uncivil behaviors,[2] includes examples of incivility for both violations of politeness norms and collective norms. More specifically, the operationalization encompasses seven key categories that, to a varying degree, have been used frequently to study incivility in user comments (e.g., Coe et al., 2014; Kalch & Naab, 2017; Muddiman & Stroud, 2017; Papacharissi, 2004). We are, therefore, optimistic that the operationalization allows answering the research questions that will be established in the following sections.

3. Consequences of incivility

Incivility in comment sections can have detrimental effects on other users as well as on democratic processes in general. According to the *Civility in America* report, a majority of the Americans believe that incivility in general leads to more discrimination (88%), less community (83%), and less political engagement (75%; Weber Shandwick, Powell Tate, & KRC Research, 2017). Experimental research has shown that reading uncivil comments can increase readers' aggressive cognitions and negative emotions (Gervais, 2015; Rösner, Winter, & Krämer, 2016), and promote stereotypic thinking about social groups (Hsueh et al., 2015). Furthermore, these comments can polarize individuals' opinions on social issues (Anderson et al., 2014). Studies have also found evidence that reading uncivil user comments makes readers communicate in a less civil manner themselves (Gervais, 2015; Hsueh et al., 2015; Ziegele, Weber, Quiring, & Breiner, 2018). Finally, these comments can deteriorate users' perception of journalistic quality (Prochazka et al., 2018) and make users refrain from participating in online discussions, which, in consequence, could inhibit the public expression of opposed, legitimate, and critical opinions (Ziegele, 2016).

On an institutional level, incivility in user comments has made one out of three German editorial departments restrict their comment sections (Meedia, 2016). Some news outlets in Germany and the U.S. have also shut down their comment sections completely (Stroud et al., 2015; Wüllner, 2015). Still, comment sections continue to exist on the Facebook sites of the media outlets (Rowe, 2015a; 2015b), where the opportunities for restrictive forms of moderation are more limited than on the websites of the news outlets (Ziegele & Jost, 2016). In the social media environment, hence, less restrictive forms of journalistic moderation, such as interactive moderation could be a solution to overcome the challenges of incivility in online discussions and to improve the civility of users' contributions in comment sections. The following section will outline these thoughts in more detail.

2 Types of personal-level incivility that were not considered include aspersions, hyperboles, words that indicate non-cooperation, and pejorative speak (Papacharissi, 2004). Types of public-level incivility that were partly considered include threats of violence (as a form of threat of democracy and threat of other individuals´ rights) and antagonistic stereotypes (which also partly cover forms such as racism and sexism; Papacharissi, 2004). Types of public-level incivility that were not considered include ideological extremity and lack of comity, lack of compromise, and nonpublic acts (Muddiman, 2017).

4. Interactive moderation of (uncivil) comments

Civility, in most theoretical approaches, is considered as an essential part of deliberation processes (e.g., Friess & Eilders, 2015). The previous section has shown that incivility can significantly threaten these deliberative processes and the deliberative outcomes of online discussions on news websites and social media platforms. Therefore, many journalistic interventions focus on countering incivility (e.g., Esau, Friess, & Eilders, 2017). This section will draw on deliberation research to evaluate and classify these incivility-related interventions and to provide a preliminary taxonomy of how journalists can interactively engage with uncivil comments to foster civil and deliberative discussions.

Many news media outlets, to reduce the number of uncivil comments and to counteract their detrimental effects, regulate their comment sections by applying various forms of *moderation*. Moderation can be defined as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" (Grimmelmann, 2015, p. 47). One form of moderation, which news media outlets can apply on their websites and partly on their Facebook sites, is *content moderation* (Wright, 2006). It includes using manual or automated filtering methods that sort out comments with pre-defined (uncivil) words before their publication (Diakopoulos & Naaman, 2011), deleting comments that are already published if they include inappropriate content, or allowing users to report these comments as abusive (Ksiazek, 2016).

These forms of content moderation, although necessary when comments violate existing laws, have been criticized for inadequately limiting citizen participation and free speech (Janssen & Kies, 2005; Wright, 2006). Put differently, content moderation can lead to an imbalance of power between journalists and users, and thereby could violate norms of deliberative processes (Friess & Eilders, 2015). For example, most forms of content moderation are non-discursive, and therefore do not adhere to the norm of reciprocity because they neither allow commenters to interactively respond to the moderators who deleted their comments, nor to learn about the reasons that led to this decision (Grimmelmann, 2015). Furthermore, many commenters use subtle forms of incivility, which moderators often do not reject, presumably because they perceive no justification for deletion or pre-moderation in these comments (Chen, 2017; Muddiman & Stroud, 2017). Finally, speaking in terms of this special issue, content moderation represents no true counter-voices because the users' statements and the moderators' objections are not publicly visible and traceable for readers.

Therefore, to foster civil and deliberative discussions, the moderation itself could benefit from living up to the standards of deliberation. For these reasons, the present study investigates a form of moderation, which previous research has coined as interactive moderation (Wright, 2006; Ziegele & Jost, 2016) or *engaged moderation* (Stroud et al., 2015). Interactive moderators participate actively and visibly in the discussions by performing a broad range of interactions with the comments of their users, such as answering questions, providing additional information, keeping discussions on track, and complimenting users for thoughtful comments (Ziegele & Jost, 2016). Particularly on social media plat-

forms such as Facebook, more and more news outlets interactively moderate the discussions of their users below their articles (e.g., Reuter, 2016).

Various studies have shown that users generally appreciate such journalistic engagement (Bergström & Wadbring, 2014; Diakopoulos & Naaman, 2011; Stroud et al., 2016). Journalists, however, are split on whether they should respond to the comments of their users (Loke, 2012). Many journalists felt that "interacting with commenters was not a journalistic value" (Nielsen, 2014, p. 480). For these journalists, it is important to maintain professional journalistic standards and avoid intruding a user space (Diakopoulos & Naaman, 2011).

This ambivalence is reflected in theoretical considerations about the effects of interactive moderation. From a deliberative perspective, such moderation could foster the development of deliberative norms and behaviors, reduce incivility, and improve the discussion atmosphere in comment sections without overly limiting free speech (Meyer & Carey, 2014; Stroud et al., 2015). Two mechanisms could account for these effects: From a social learning perspective (Bandura, 1977), users could perceive the behavior of moderators who perform deliberative interactions as examples and adapt their behavior according to these examples. From the perspective of social norms, the presence of interactive moderators could constitute an external normative influence and increase users' awareness that "others are watching" (Stroud et al., 2015, p. 191). Such an awareness can encourage conformity, particularly in the case of public behaviors (Lapinski & Rimal, 2005). On the other hand, when moderators act in too obtrusive ways (i.e., telling users how to behave), the opposite effect might occur; that is, users could show reactance to a seemingly inappropriate journalistic intrusion of "their" space. In fact, a survey of commenters and comment readers showed that while these users supported that journalists clarify factual questions in comment sections, most of them disagreed that journalists should direct the discussions (Stroud et al., 2016).

Users' different approval of various forms of interactive moderation suggests that the way journalists respond to different comments will impact the effects of moderation. To date, however, no systematic overviews of different journalistic moderation styles have been provided. In Table 1, we therefore differentiate between four journalistic responses to user comments. This differentiation, although certainly not comprehensive, builds on models of deliberation (e.g., Friess & Eilders, 2015), behavioral psychology (e.g., Cheng, Danescu-Niculescu-Mizil, & Leskovec, 2014), the roles of (interactive) moderators in online discussions (Wright, 2006), various task descriptions of (interactive) moderators (e.g., Grimmelmann, 2015), and studies on interactive moderation of user comments (Stroud et al., 2015; Ziegele & Jost, 2016). Additionally, with these theoretical considerations in mind, we conducted a qualitative analysis of 100 moderation comments that were randomly selected from the corpus of comments used for this study.[3] This combination of deductive and inductive processes resulted in a taxonomy

3 This analysis involved elements from qualitative content analysis (Mayring, 2000) and discourse analysis (Herring, 2004). The procedure was adopted from Ziegele (2016). Please refer to this publication for further information.

that classifies journalistic responses to (uncivil) user comments according to their deliberativeness and the kind of behavioral sanction (Table 1).

Table 1. Taxonomy of interactive journalistic moderation

		Kind of sanction	
		Reward	**Punishment**
Deliberativeness	**Deliberative**	**Discursive moderation** Factually engaging with comments; providing additional information; clarifying questions; adding arguments.	**Regulative moderation** Factually complaining about comments; asking users to behave more civilly; pointing to violations of predefined rules.
	Non-deliberative	**Sociable moderation** Informally complimenting comments; creating an informal and pleasant discussion atmosphere.	**Confrontational moderation** Offensively attacking comments; using irony/sarcasm to expose comments to ridicule.

Basically, a moderator can either respond to a comment with a reward (positive feedback) or a punishment (negative feedback; cf. Cheng et al., 2014). According to the operant conditioning framework from behavioral psychology (Skinner, 1938), positive feedback can motivate users to contribute high-quality comments, and punishment can make users contribute fewer low-quality comments (Cheng et al., 2014). Both types of responses can be written in a deliberative or a non-deliberative way. Deliberative responses hint or adhere to the norms of deliberative discussions, which are civility, mutual respect, rationality, and constructiveness (Friess & Eilders, 2015). Deliberative rewards (*discursive moderation*) occur when moderators show respect to comments and their authors by performing deliberative interactions, such as clarifying questions, providing additional information, or adding arguments. An example from our data reads as follows:

> *User:* Can somebody explain why his [a presumed terrorist's] face is now anonymized? It's not like we wouldn't know him anyway…
>
> *Moderator:* Why he is now being made anonymous: The terror suspect had been arrested since Monday. From then on, the same applies to him as to all offenders who end up in court: His face is pixelated and his name anonymized.

Moderators responding to comments with deliberative punishments (*regulative moderation*) point out the deliberative norms that were violated by the respective comment, thereby informing the commenters about the low quality of their comments and directly aiming at improving the quality of their future contributions. As shown in the following example from our data, such a punishment is often combined with an announcement that the comment will be deleted, or the user blocked.

> *User:* This has been a method since 1990. It [Chemnitz] is a right-wing radical region with right-wing radical state institutions. We all know where this will end. The whole area remains the shame of Germany.
>
> *Moderator:* Please stay objective—thank you! We will delete insults.

Our qualitative pre-study, as well as previous research on community moderation (e.g., Wright, 2006), has shown that, particularly on social media platforms, moderators do not always respond to commenters in a deliberative manner. Therefore, two other moderation styles in the taxonomy include non-deliberative rewards or punishments. Regarding non-deliberative rewards (*sociable moderation*), community research has recommended moderators to engage with their community in a sociable way (Kraut, Resnick, & Kiesler, 2011), for example by greeting community members, appreciating their comments, or by creating an informal and sociable discussion atmosphere with harmless jokes and small talk. One example in our data reads as follows:

> *User:* @n-tv: Can you give me your definitions of "refugee" and "terrorist"? And if you're really smart, maybe you can tell me what you did wrong.
>
> *Moderator:* Can we use the telephone joker? :-)

Deliberation in the Habermasian sense focuses on the rational exchange of arguments (Habermas, 1996) and does not include such informal and community-oriented responses. This also applies to the second non-deliberative moderation style, non-deliberative punishments (*confrontational moderation*): Some news outlets, such as *Die Welt*, sometimes attack users who write inadequate comments in combative ways, often by using sarcasm or cynicism, "to hold them up a mirror and reveal the inappropriateness of their behavior" (Ziegele & Jost, 2016, p. 6). Such moderation itself is close to incivility and might therefore be perceived as non-deliberative (ibid.):

> *User:* All politicians are crappy, no matter who you elect!! [The original German post contained many spelling errors that have not been included in the translation].
>
> *Moderator:* Don't forget to read the dictionary before voting!!! [The original German post contained many spelling errors that have not been included in the translation]

Having described these different styles of interactive moderation, the questions remain (a) which comments are related to increased interactive moderation; and (b) whether interactive moderation and different moderation styles are successful in civilizing the following reply comments. Answers to these questions could contribute to developing a realistic assessment of the patterns of journalistic counter-voices in comment sections and can help designing adequate moderation strategies. However, little is known about these patterns of interactive moderation and about its potential effects. Previous research has investigated journalistic moderation in general (i.e., not differentiating between content moderation or interactive moderation; Wise, Hamman, & Thorson 2006), analyzed interactive moderation of civil comments (Stroud et al., 2015), or is limited to experimental settings (Ziegele & Jost, 2016).

As this special issue focuses on problematic communication behavior, we will investigate these questions using the example of uncivil comments. Due to the detrimental effects of uncivil comments reported in the previous section, it could be particularly important that interactive moderators respond to such comments to show the respective commenters that incivility is not a tolerated behavior, and to provide comment readers a more differentiated perspective on the respective issues.

However, the previous section has shown that there are different levels of incivility, namely personal-level and public-level incivility. Although the use of personal-level and public-level incivility both violate social norms, moderators have limited capacities and could therefore focus on easily recognizable forms of incivility, such as name-calling or profanity (Muddiman & Stroud, 2017). Additionally, moderators could generally refrain from responding to uncivil comments and focus on engaging with comments that already adhere to deliberative norms. To date, no research has investigated the characteristics of comments that are associated with increased interactive moderation activities. Therefore, we pose our first research question:

RQ1: Is incivility in user comments associated with increased interactive journalistic moderation (RQ1a) and, if yes, is this pattern consistent across different levels of incivility (RQ1b)?

Second, research has not yet investigated how journalistic moderators respond to uncivil comments. The four different moderation styles described above are predominantly derived from theoretical investigations and a qualitative study. Although one might assume that some styles, such as a sociable moderation, are used less frequently when responding to uncivil comments, no quantitative research has investigated this assumption to date. An assessment of the styles that moderators use to respond to uncivil comments in general and to different levels of incivility in particular can shed light on if and how journalists are trying to enforce deliberative norms in comment sections. Additionally, such an assessment is a prerequisite for analyzing the potential effects of different moderation styles of initial comments on the quality of users' subsequent reply comments. Therefore, we ask a second research question:

RQ2: What is the distribution of sociable, discursive, confrontational, and regulative moderation in response to uncivil user comments in general (RQ2a) and in response to different levels of incivility (RQ2b)?

Finally, journalists responding to uncivil and civil comments likely aim at fostering or preserving a civil and high-quality discourse. For civil comments, a field experiment conducted in the U.S. showed that discursive moderation of these comments further increased the deliberativeness of the discussions; users wrote more civil comments and provided more evidence when identifiable journalists participated in the discussions and answered questions, for example (Stroud et al., 2015). Yet, these effects did not occur when the moderators were anonymous, that is, when the news outlet itself was displayed as the author of the moderation comments. On social media platforms, however, the latter is much more common; usually, news outlets identify as authors of moderation comments. Additionally, the study investigated only one moderation style. Regarding uncivil comments, results of a lab experiment have demonstrated that a regulative moderation of these comments (i.e., politely asking users to discuss more civilly) made readers of the discussions perceive a more deliberative discussion atmosphere. No such effect occurred when the moderator responded to the uncivil commenter using a confrontational moderation style (Ziegele & Jost, 2016). Another study showed that punishing users for their comments even lowered the quality of their future

comments (Cheng et al., 2014). Rewards neither lowered nor increased comment quality. Due to these ambivalent findings, we ask whether different moderation styles of initial comments will increase or decrease the presence of incivility in the comments that reply to these initial comments:

RQ3: How do regulative, discursive, sociable, and confrontational moderation styles of initial comments relate to the presence of incivility in the comments that reply to these initial comments?

5. Method

The current study is based on a quantitative content analysis of Facebook posts of 15 German news media outlets. The sample covered a broad range of formats and it included public service and private media, nation-wide and regional organizations, and both conservative and liberal outlets. Table 2 provides an overview of the news media outlets in the sample. A media outlet was included a) if it was among the more than 2,000 German media outlets covered in a comprehensive list (Pertsch, 2016), b) if it had more than 15,000 followers on Facebook, c) if it was primarily a *news* media outlet, and d) if the news outlet had engaged in interactive moderation during February 2016. To identify the news outlets that engaged in interactive moderation, a student assistant scrolled through the Facebook posts of each outlet with more than 15,000 followers on the list in February 2016. As soon as the student assistant identified a moderation comment below these posts, the outlet was included in the sample.

To reduce a potential sampling bias caused by specific events and periods (e.g., holidays, terrorist attacks, scandals, etc.), the period of investigation was split: In February and October 2016, all articles (N = 10,081) and comments (N = 2,112,897) published on the news outlets' Facebook pages were crawled with the help of the tool *netvizz*. Within this corpus, *netvizz* classified the authors of the moderation comments as "pageowner," which allowed the researchers to quickly identify these comments. Based on Facebook's nested comment structure, *netvizz* further classified each comment as an *initial comment* (a top-level comment, respectively) or a *reply comment* (i.e., a comment replying to an initial comment). This information was used to draw a stratified random sample in three steps: First, on the level of reply comments, we randomly selected up to 100 moderation comments per media outlet and month[4]. In the second step, using automated ID matching, the corresponding initial comments to which these moderation comments reply were collected. These comments, therefore, constituted the sample of initial comments *with* moderation. Finally, we randomly selected up to 100 initial comments from the same discussion thread that, on the level of reply comments, had received no moderation comment. These comments constituted the sample of initial comments *without* moderation.

4 100 comments per category, news outlet, and month was the maximum we could expect of our student coders, who received their course credit in exchange for a maximum of 15 hours of coding work.

Table 2. Overview of the news media outlets in the sample

News outlet	Followers on FB (9/18)	Format	Distribution	Financing	Political Orientation
ARD Tagesschau	1,584,969	TV	NAT	PUB	Moderately left-liberal
Spiegel Online	1,553,499	NM	NAT	PRI	Left-liberal
Die Welt	978,227	NP	NAT	PRI	Conservative
n-tv	917,176	TV	NAT	PRI	n/a
ZDF heute	875,774	TV	NAT	PUB	Moderately right-liberal
Süddeutsche Zeitung	741,387	NP	NAT	PRI	Center-left
Frankfurter Allgemeine	520,427	NP	NAT	PRI	Center-right
Berliner Morgenpost	258,097	NP	NAT	PRI	Conservative
BR24	219,186	RA	REG	PUB	n/a
RTL2 News	201,966	TV	NAT	PRI	n/a
Deutschlandfunk	179,455	RA	NAT	PUB	Center-left
Tagesspiegel	146,887	NP	NAT	PRI	Liberal-conservative
Hannoversche Allgemeine	94,258	NP	REG	PRI	n/a
Krautreporter	91,280	ON	W	PRI	n/a
HR-info	28,691	RA	REG	PUB	n/a

Notes. TV = TV news, NM = news magazine, NP = Newspaper, RA = Radio news, ON = Online news site, NAT = National, REG = Regional, W = Worldwide, PUB = Public, PRI = Private, Evaluation of political orientation based on Eilders (2002), euro I topics (2018a;b), and Maurer & Reinemann (2006). When no political orientation is reported, the information was not available for the corresponding news outlet.

Theoretically, for each media outlet, a maximum of 600 comments for the two months could have been coded (200 initial comments with moderation, 200 initial comments without moderation, and 200 moderation comments). However, during the sample period, four media outlets wrote less than 200 moderation comments (see Results section and Table 3 for details). Additionally, some media outlets frequently wrote multiple reply comments in a single thread. Within these threads, only the first moderation comment was included in the sample. In sum, 1,656 initial user comments with moderation were coded. To keep the groups roughly equal regarding their size, we intended to code 1,656 initial user comments without moderation. However, 106 of these comments had to be excluded because they did not include any text and could not be accessed on Facebook anymore. Therefore, only 1,550 initial user comments without moderation were coded. All comments were posted under 1,670 news articles. For the 1,656 initial comments with moderation, we also coded the characteristics of the respective 1,656 moderation comments that replied to these comments. By comparing moderated and unmoderated user comments, we identified the characteristics of the comments (i.e., incivility) that were related to increased levels of interactive moderation (RQ1) as well as to different moderation styles (RQ2).

Another aim of this study was to investigate the potential effects of the different styles of the moderation comments that reply to initial comments on the level of incivility in users' subsequent reply comments (RQ3). Therefore, up to six fur-

ther reply comments to each initial comment—both moderated and unmoderated ones—were coded (n = 4,901). This was possible because Facebook automatically attaches reply comments to the related initial comments. Consequently, in the dataset, each reply comment included the unique ID of the related initial comment. The six reply comments were selected in the chronological order they responded to the initial comment. We did not measure whether the author of the reply comment responded to the initial comment, to a moderation reply comment, or to another reply comment. Still, for each reply comment to a moderated initial comment, we coded whether it was posted before or after the moderation comment (K-α = 1, PA = 1). For some analyses, only reply comments that were posted *after* the moderation comment were selected (see Results section). In sum, 9,763 user and moderation comments were coded (initial and reply comments).

Table 3. Distribution of moderation comments across the 15 news media outlets in the corpus

	Total comments	Non-moderation comments (all)		Moderation comments (replies)		Sampled moderation comments (replies) [a)]
	n	*n*	%	*n*	%	*N*
HR-info	8,748	8,378	95.77	354	4.05	121
BR24	28,114	27,302	97.11	717	2.55	141
Die Welt	407,425	402,315	98.75	4,990	1.23	193
Krautreporter	8,879	8,765	98.72	105	1.18	87
Tagesspiegel	33,495	33,215	99.16	175	0.52	86
Hannoversche Allgemeine	37,930	37,724	99.46	194	0.51	121
Frankfurter Allgemeine	141,756	141,040	99.49	424	0.30	114
Berliner Morgenpost	53,401	53,242	99.70	99	0.19	82
Sueddeutsche Zeitung	199,347	198,901	99.78	364	0.18	140
n-tv	214,220	213,984	99.89	217	0.10	166
RTLII News	19,926	19,896	99.85	20	0.10	18
Spiegel Online	426,905	426,475	99.90	385	0.09	163
ARD Tagesschau	309,093	308,472	99.80	272	0.09	98
Deutschlandfunk	60,972	60,910	99.90	51	0.08	38
ZDF	162,686	162,457	99.86	130	0.08	88
n	2,112,897	2,103,076	99.54	8,497	0.40	1,656

Notes. [a)] Difference to 200 comments (2 months * 100 comments) caused 1) by insufficient number of moderation reply comments or 2) by multiple moderation reply comments in the same thread (then, only the first moderation reply comment was coded).

For all selected *initial comments*, 52 trained undergraduate students coded various types of personal-level incivility and public-level incivility. The coding scheme was based on previous categorizations of incivility (e.g., Coe et al., 2014; Papacharissi, 2004). It included the following types:

(a) Insults, that is, comments including name-calling (prevalence: 12%, Krippendorff's-α [K-α] = .52, percent agreement [PA] = .76);
(b) profanity, that is, the presence of vulgar or obscene language (prevalence: 5%, K-α = .69, PA = .92);
(c) accusations of lying (prevalence: 7%, K-α = .68, PA = .88);
(d) threats of violence, that is, announcing aggressive action against a target (1%, K-α = .48, PA = .89);
(e) negative stereotypes, that is, overgeneralized disparaging statements about social groups and categories (prevalence: 7%, K-α = .42, PA = .71);
(f) sarcasm and cynicism (prevalence: 9%, K-α = .39, PA = .69);
(g) and "screaming", that is, the extensive usage of capital letters (prevalence: 7%, K-α = .75, PA = .95)[5].

As the K-α coefficients were substantially below the acceptable threshold of .67 (Krippendorff, 2004) for four of the seven types of incivility,[6] the respective categories had to be summarized into two indices: Summing up the values for profanity, insults, sarcasm, and screaming into a *personal-level incivility index* increased the reliability to an almost satisfactory level (K-α = .65, PA = .83). The values for negative stereotypes, accusations of lying, and threats of violence were summed up into a *public-level incivility index*, which also had higher reliability scores than the single categories (K-α = .67, PA = .86). These indices also correspond with the differentiation between personal-level incivility and public-level incivility in the theory section. Therefore, these indices were used for all subsequent analyses.

Furthermore, the students coded the moderation style of each journalistic moderation comment. As the styles are not mutually exclusive (e.g., a moderator can criticize a user's behavior but still engage in a discussion with her or him), we allowed the students to code up to three styles in three categorical variables (moderation style 1, moderation style 2, moderation style 3). For each variable, students could code one of the four styles described in Table 1 (1 = "discursive moderation", 2 = "regulative moderation", 3 = "sociable moderation", 4 = "confrontational moderation"). For reliability testing and all other analyses, these three variables were recoded into four dummy variables. Each dummy variable contained the information whether one of the four moderation styles was present (0 = "not present", 1 = "present"). Reliability scores were satisfying for discursive moderation (K-α = .68, PA = .85), regulative moderation (K-α = .73, PA = .92), and sociable moderation (K-α = .82, PA = .93). For confrontational moderation, which occurred quite infrequently, Krippendorff's α was slightly below the acceptable value, but the percent agreement was satisfactory (K-α = .65, PA = .86).

5 The reliability scores are based on a randomly selected sample of 15 news articles and 75 related user comments that were coded by all students.

6 It is possible that it is particularly challenging to detect incivility and hate speech in German language because other German researchers have reported even lower coefficients (e.g., Ross et al., 2016). Another reason why our coefficients were rather low is that the comments were coded by more than 50 student coders. Other research (including the study by Coe et al., 2014) used only three to five student assistants to code the data. Achieving agreement between all coders gets more difficult the more coders there are.

1. The discursive moderation (prevalence: 60% of all moderation comments) can be seen as a deliberative reward; here, moderators respond to a comment with own substantial arguments, answer questions, provide additional information, or stimulate conversation among discussants and, thereby, promote deliberative exchange.
2. The regulative moderation (prevalence: 12%) can be described as a deliberative punishment. It aims at enforcing predefined rules or keeping discussions on topic. Therefore, moderators admonish users to behave in the desired manner when they violate the norms of civility or moderators mediate conflicts among users.
3. The sociable moderation (prevalence: 32%) functions as a non-deliberative reward. Moderators applying this style foster a pleasant discussion atmosphere by making harmless jokes, engaging in small talk, or by applauding good comments. In doing so, the sociable moderation style aims at impeding incivility in a subtle way.
4. The confrontational moderation (prevalence: 9%) pursues the same goal in a non-deliberative manner (i.e., non-deliberative punishment). In contrast to the regulative style, confrontational moderation aims at silencing the authors of inappropriate comments by using ironic, sarcastic or confrontational statements.

Finally, the students coded the presence (or absence) of incivility in the first six reply comments to the initial comments with and without moderation (n = 4,901). Owing to the limited resources of the project, a quick dichotomous measure of incivility had to be used to code the presence of uncivil elements in the reply comments. That is, the coders decided for each of the maximum of six reply comments whether they contained any of the above-mentioned types of personal-level or public-level incivility (prevalence: 14% of all reply comments, K-α = .83, PA = .90).

6. Results

Twenty-five percent of the initial user comments in the sample contained at least one type of personal-level incivility. Violations of collective norms occurred less frequently, with 14 percent of all initial comments containing at least one type of public-level incivility. Both the prevalence of personal-level and public-level incivility differed heavily between the news media outlets. The prevalence of personal-level incivility ranged between 16 percent on the Facebook page of the *Berliner Morgenpost* and 35 percent on the Facebook page of *Deutschlandfunk*. The comments on the Facebook page of the *Berliner Morgenpost* also showed the lowest share of public-level incivility (7%), while the highest share of public-level incivility was observed in the comments posted on the pages of *Deutschlandfunk* and *BR24* (19%).

Journalistic engagement was rarely observable in the discussions on the Facebook pages of the media outlets investigated (see Table 3): only 8,497 of the 2,112,897 comments in our corpus stemmed from the media outlets themselves

and replied to other comments (0.40%). Relatively speaking, the radio channel *hr-info* moderated the highest share of user comments on its Facebook site (4.05%, 354 of 8,748 comments), followed by the radio channel *BR24* (2.55%, 717 of 28,114 comments), and the newspaper *Welt.de* (1.23%, 4,990 of 407,425 comments). The lowest shares of moderation comments were visible on the Facebook site of the public broadcaster *ZDF* (0.08%, 130 of 162,686 comments) and the radio channel *Deutschlandfunk* (0.08%, 51 of 60,972 comments).

Table 4. Logistic regression of the presence of interactive moderation on the presence of incivility (RQ1a) and different levels of incivility (RQ1b)

	Model 1			Model 2		
	B	*SE*	*Odds*	*B*	*SE*	*Odds*
Presence of incivility	0.33*	.08	1.39	-	-	-
Personal-level incivility	-		-	0.12	.09	1.13
Public-level incivility	-		-	0.39***	.11	1.48
Controls				0.56	.18	1.75
n-tv	0.55**	.17	1.73	0.44**	.18	1.55
Sueddeutsche Zeitung	0.44*	.18	1.55	0.01*	.18	1.01
Frankfurter Allgemeine Zeitung	0.00	.18	1.00	0.11	.20	1.12
Berliner Morgenpost	0.10	.20	1.11	0.41	.27	1.51
Deutschlandfunk	0.41	.27	1.50	0.58	.18	1.78
BR24	0.58*	.18	1.78	0.82**	.20	2.27
HR-info	0.81***	.20	2.24	0.77***	.21	2.15
Krautreporter	0.77***	.21	2.16	0.46***	.17	1.58
Die Welt	0.45**	.17	1.57	0.03**	.19	1.03
ZDF heute	0.03	.19	1.03	0.74	.19	2.10
Hannoversche Allgemeine	0.74***	.19	2.09	0.01***	.19	1.01
Tagesspiegel	-0.01	.19	1.00	0.44	.17	1.55
Spiegel Online	0.43*	.17	1.54	0.68*	.38	1.97
RTLII News	0.66	.38	1.93	0.39	.11	1.48
ARD Tagesschau (Ref Cat.)	1			1		
R^2 (Nagelkerke)	.032			.035		
N	3,206			3,206		

Notes. Dependent variable: Presence of interactive moderation (*1* = "yes"). * $p < .05$, ** $p < .01$, *** $p < .001$

The first research question asked whether uncivil comments are related to increased levels of interactive journalistic moderation compared to other comments (*RQ1*). In other words, we aimed at explaining whether an initial comment will be moderated (yes/no; DV) depending on the presence and specific levels of incivility (personal-level and public-level) in this comment (IV). As described above, the discussions on the sites of the various media outlets in the sample differed regarding their overall share of incivility. Furthermore, the media outlets also differed regarding their likelihood to moderate and regarding their moderation style (Table 3). To control for these effects, the different news outlets were added to the

analysis models. A stepwise logistic regression[7] initially showed a positive correlation between the presence of incivility and the likelihood of interactive journalistic moderation (RQ1a). This effect prevailed after adding the control variables to the model (B = 0.33, $Odds$ = 1.39, $p < .05$). As expected, the media outlets themselves also differed regarding their likelihood to respond to comments (Table 4, Model 1).

We then investigated whether personal-level incivility (e.g., name-calling and shouting) and public-level incivility (e.g., antagonistic stereotypes and threats) both relate to increased interactive moderation (*RQ1b*). A second logistic regression (Table 4, Model 2) showed that, after including the control variables, only public-level incivility in initial comments was related to increased interactive moderation (B = 0.39, $Odds$ = 1.48, $p < .001$).

Table 5. Moderation styles of uncivil comments and of different levels of incivility

Presence of incivility / Moderation styles	No incivility	Incivility	Personal-level incivility	Public-level incivility
	%	%	%	%
Regulative moderation	6	18	16	15
Discursive moderation	48	61	46	52
Confrontational moderation	6	16	13	14
Sociable moderation	41	31	26	19
n	1,244	688	541	329
(%)	(100)	(100)	(100)	(100)

Notes. For each moderation comment, the coders were allowed to code up to three moderation styles. Therefore, the percentages describe the share of the respective moderation style on all moderation styles coded in the moderation comments replying to the respective category of incivility.

RQ2 asked about the distribution of the styles journalists use when responding to uncivil comments in general (*RQ2a*) and to different levels of incivility (*RQ2b*). Table 5 shows that when journalists responded to civil comments, they primarily used a discursive style (48%), followed by a sociable style (41%), a confrontational style (6%), and a regulative style (6%). For comments that contained any level of incivility, the share of regulative moderation tripled (18%), while the share of sociable moderation declined (31%). The share of both confrontational moderation (16%) and discursive moderation (61%) increased. Thus, moderators consistently engaged with uncivil comments more often using negative sanctions and less often using non-deliberative rewards. No fundamental differences were found regarding how moderators responded to different levels of incivility; however, compared to personal-level incivility, moderators replied to public-level incivility less often using a sociable style, and more often using a discursive style.

Finally, we investigated how different journalistic moderation styles of initial comments relate to the level of incivility in users' reply comments to these initial

7 We did not conduct multilevel analyses with the media outlets as level-2 variables (group level) and the comments as level-1 variables because the small number of 15 cases at the group level would likely produce biased estimates (e.g., Maas & Hox, 2005).

comments (*RQ3*). For this purpose, the incivility (i.e., the presence of at least one form of incivility) of the first six comments that replied to the respective initial comments was coded. Both moderated and non-moderated initial comments were included in the analysis. For the moderated initial comments, only those reply comments were selected that were posted after a moderation comment. A total of 1,464 initial comments and 3,943 reply comments were included in the analysis (see Table 6).

Table 6. Multilevel hierarchical logistic regression of the incivility of reply comments (level 1) on initial comment characteristics and moderation characteristics (level 2)

	Incivility of a reply comment		
	B	*SE*	*t*-ratio
Characteristics of the initial comment			
Personal-level incivility	0.89***	0.14	7.57
Public-level incivility	0.80***	0.12	5.76
Characteristics of the moderation			
Discursive moderation	0.15	0.12	1.27
Regulative moderation	0.41*	0.17	2.44
Confrontational moderation	0.20	0.18	1.12
Sociable moderation	-0.39**	0.14	-2.87
Controls			
n-tv	0.01	0.27	0.05
Sueddeutsche Zeitung	-0.12	0.28	-0.42
Frankfurter Allgemeine Zeitung	-0.41	0.33	-1.26
Berliner Morgenpost	-0.91*	0.40	-2.28
Deutschlandfunk	0.10	0.43	0.23
BR24	-0.13	0.27	-0.48
HR-info	-0.17	0.28	-0.59
Krautreporter	-1.19*	0.33	-3.58
Die Welt	-0.20	0.26	-0.77
ZDF heute	0.13	0.31	0.42
Hannoversche Allgemeine	-0.47	0.31	-1.53
Tagesspiegel	-0.24	0.35	-0.68
Spiegel Online	-0.14	0.27	-0.53
RTLII News	-0.67	0.54	-1.24
ARD Tagesschau (Ref Cat.)	1	1	1
n (initial comments)		1,464	
n (reply comments)		3,943	

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

All initial comments, moderation comments, and reply comments are nested in the discussions below the posts. The dependent variable on level 1 was the presence of incivility in a reply comment. This was coded dichotomously (0 = "not

present", 1 = "present"). Therefore, a multilevel logistic regression model was applied (Hox, 2010). We assumed that both the level of incivility of initial user comments and the presence and styles of moderation comments (independent variables on level 2) would relate to the level of incivility in the related reply comments. Therefore, the personal-level and public-level incivility variables were added to the regression model ("characteristics of initial comments" in Table 6). To reveal if the four moderation style variables were associated with the presence of incivility in the subsequent reply comments, these styles were then added to the model ("characteristics of the moderation" in Table 6). Finally, the different news media outlets were included as control variables ("controls" in Table 6).[8]

Results show that the comments replying to the related initial comments were more uncivil when the initial comments contained personal-level incivility (B = 0.89, SE = 0.14, p < .001) and public-level incivility (B = 0.80, SE = 0.12, p < .001). Regarding moderation styles, when journalists told uncivil commenters to discuss more civilly (*regulative moderation*), this was related to even more uncivil reply comments (B = 0.41, SE = 0.17, p < .05). In contrast, a *sociable moderation style* reduced the incivility of the subsequent reply comments (B = -0.39, SE = 0.14, p < .01). Neither a confrontational nor a discursive moderation of initial comments was associated with higher or lower levels of incivility in the respective reply comments.

7. Discussion

Politicians, scientists, and journalists are searching for strategies to deal with incivility in public user discussions and its potentially harmful effects on other users. Interactive moderation has been considered as a particularly promising strategy. Using a quantitative content analysis, the current study shows that news outlets indeed respond to comments that contain incivility, that they primarily do so in a discursive manner, and that the style of journalistic responses to initial comments relates to different levels of incivility in the related reply comments. More specifically, the findings imply that journalists responding to comments in a sociable manner decrease the level of incivility in the subsequent reply comments. This finding coincides with research on online communities that recommends moderators to maintain a respectful and friendly tone even in challenging situations (Kraut et al., 2011). It is also supported by theories such as the *verbal-person-centered theory of social supportive outcomes* (Bodie & Burleson, 2008; Chen, Riedl, & Huang, 2018), which has claimed that high-person centered responses, which recognize and acknowledge a person's (negative) feelings, are more effective than other messages. On a preliminary level, the results of the current study suggest that showing respect and empathy for the feelings (not necessarily the positions) even of uncivil commenters could help to improve the quality of the subsequent reply comments.

8 See footnote 7 for the reasons why we did not include the media outlets on a separate level.

In contrast, the reply comments were even more uncivil when journalists responded to the related initial comments in a regulative manner. This finding is in line with previous research on online communities, which found that users who received negative feedback by other users wrote more low-quality comments in the future (Cheng et al. 2014). Such behavior can be explained by reactance theory (for an overview, see Miron & Brehm, 2006): users could feel illegitimately patronized by regulative moderation and answer even more uncivil. Moreover, moderators' capacities to impose sanctions are limited. Thus, assaulted users might tend to seek relief from their negative emotions by expressing their anger, knowing that the moderator will probably not intervene again.

Furthermore, our results show that most news outlets did not consistently engage with different levels of incivility. In fact, only the use of public-level incivility in comments, such as negative stereotypes, accusations of lying, and threats, was related to an increased level of interactive moderation. In contrast, the use of personal-level incivility in comments, such as name-calling and shouting, neither increased nor decreased the likelihood of interactive moderation. At the first glance, this result suggests that media outlets consistently perform their roles in society as generators and preservers of civil *public* discourse: public-level incivility threatens collective norms of public discussion and deliberation (Muddiman, 2017; Papacharissi, 2004). At the same time, many users communicate this kind of incivility in a quite subtle way – such as applying latent stereotypes. Therefore, newsrooms often cannot simply delete comments that include public-level incivility. Interactively responding to these comments therefore is particularly important for institutions that constitute public discourse and hold a certain degree of public responsibility. Regarding impoliteness or personal-level incivility, newsrooms might assume that these forms of speech—at least to a certain degree—do not necessarily undermine collective norms and traditions of democracy and is therefore not needed to be censured (Papacharissi, 2004). In the long term, such an assessment, however, might fall short, because studies have found that citizens perceive some manifestations of personal-level incivility as even more severe than some types of public-level incivility (Muddiman, 2017). Consequently, a lack of (interactive) moderation of these forms of personal-level incivility in user comments could likely discourage users from maintaining a positive attitude towards the news media outlet hosting these discussions or from participating in comment sections (Ziegele, 2016).

In sum, it is challenging for journalists to enforce norms of civility in online discussions via interactive moderation. Still, it is possible that other forms of interactive moderation, such as regulative or discursive moderation, increase other qualities of the subsequent reply comments, such as the use of arguments or the provision of evidence (e.g., Stroud et al., 2015). Another reason for the unexpected (non-) effects of regulative and discursive moderation may also be that the news outlets we analyzed moderated comments anonymously, that is, they did not identify as individuals but rather as the news brand as a whole. Previous research has found that interactive moderation is particularly successful when identifiable journalists engage in the discussions (Stroud et al., 2015).

After all, even if interactive moderation strategies may not be able to tame incivility completely, these strategies are, at least, a promising addition to noninteractive forms of moderation. Following a deliberative approach, interactive moderation is more transparent, open, reciprocal, and less restrictive than content moderation. Additionally, although some interactive moderation styles might not be successful to enforce norms of deliberative discussions in the short term, it might be successful to do so in the long term. Although users, in their immediate reply comments, seem to respond to some interactive moderation styles in an even more uncivil manner, they might adjust their behavior after some time (see next section).

7.1. Limitations

The findings of the current study should be interpreted only in light of several methodological limitations. First, our operationalization of incivility did not reflect the full spectrum of possible uncivil behaviors. Although we included seven types of behaviors that have been used frequently in previous studies, other types were not considered. Still, we perceive that many of the behaviors we did not measure individually partially overlap with those we measured. For example, we did not decidedly measure pejorative language (Coe et al., 2014), but this category is likely reflected by antagonistic stereotypes and insults, which also use pejorative words or grammatical forms. Nevertheless, future research should aim at measuring the possible types of incivility more comprehensively.

Second, the reliability of our incivility measures was not consistently satisfactory. We already discussed the problem that incivility lies in the eye of the beholder, and our reliability scores underline this interpretation. Although the researchers were able to increase these scores by summing up the single types of incivility into two indices—personal-level incivility and public-level incivility—the results should be interpreted with caution, and future research is needed to corroborate or refute our conclusions. This limitation also applies to the differentiated measures of incivility in initial and reply comments. For the reply comments, we could only use an overall measure for the presence or absence of incivility. This measure sufficed to distinguish civil from uncivil reply comments, and therefore to answer the research question whether different moderation styles relate to the presence of incivility in the subsequent reply comments. Still, this study could not determine whether there was more or less personal-level incivility *or* public-level incivility (or both) in the reply comments. Future research, therefore, should address this question by using differentiated measures of the incivility construct.

Third, the current study focused on a relatively short time span of two months. Therefore, potential long-term effects of journalistic moderation could not be measured. The study also only examined the immediate relations between interactive journalistic moderation of initial comments and the level of incivility of the related reply comments of the subsequent discussion. It did not investigate whether these journalistic counter-voices generally were related to the quality of the discussions overall and/or to the quality of subsequent discussions. Investigating this is important because, in the long term, especially discursive and sociable in-

teractive moderation could cause a general positive shift in the quality of users' discussions and contributions—that is, users could benefit from the information introduced by moderators applying a discursive style and, thus, write more substantial comments in the long-run. Moreover, following a social learning perspective (Bandura, 1977), users could also learn from and finally adapt the behavior of moderators who perform deliberative interactions. In addition, a sociable moderation could strengthen the ties between the media outlets and their audiences. Ideally, users then feel responsible and actively help to further improve the discussion climate. Future studies should take this into account and examine the immediate *and* long-term effects of interactive journalistic moderation on the quality of discussions overall.

Fourth, even on Facebook, many news outlets apply filtering methods to detect harmful comments and delete them if they include inappropriate content or if users report these comments as abusive (Ksiazek, 2016). As a consequence, the current study could only examine the comments that were still visible at the time the articles and comments were crawled. It is important to take into consideration that some highly uncivil comments, which were deleted after some time, might already have influenced users' discussions in the comment sections. Future studies should therefore try to include filtered and deleted comments in their analyses. Moreover, it could be beneficial to investigate the practice of combining forms of content moderation and interactive moderation (e.g., publicly explaining the reason why a comment was deleted).

Fifth, this study focused on strategies to reduce incivility and its negative effects. Furthermore, it only examined uncivil user comments and did not consider potential effects of interactive moderation on other comments. To analyze the quality of online discussions, future studies should also take other aspects of deliberation into account, such as reciprocity and constructiveness, since in some comments, incivility and other elements of deliberation coexist (Chen, 2017). Future research might also shift the focus to the analysis of positive effects of journalistic intervention on readers' knowledge and attitudes, and the effects of interactive journalistic moderation of civil comments.

Finally, using a content analysis, the current study cannot draw causal inferences regarding the effects of interactive moderation. Future experimental research is needed to corroborate or refute our findings regarding the relationships between interactive journalistic moderation and incivility in comment sections.

7.2. Implications and future research

Moderation as a strategy to prevent and counter incivility requires extensive personnel resources. Therefore, developing (semi-)automated tools that help news media outlets identify and respond to uncivil comments more systematically is an important task (e.g., Goodman, Cherubini, & Waldhorn, 2013). (Semi-)automated tools could assist moderators in detecting more obvious forms of personal-level incivility, such as insults or vulgarity. Human moderators could then spend more time on detecting and engaging with subtle but harmful forms of public-level incivility, such as the assignment of stereotypes or threats of individual

rights. For this purpose, a shared moral code across different news outlets of what is considered as uncivil could be helpful. Additionally, moderators of comment sections need to be trained in detecting different types of incivility in comment sections. Such training could also include the use of different styles of moderation. In fact, the results of the current study suggest that community managers should develop a mixed-mode moderation strategy to engage uncivil commenters in a discursive yet respectful and sometimes even sociable way. Even then, however, interactive moderation cannot entirely solve the problem of online incivility. When uncivil user comments violate existing laws, other strategies, such as automatic deletion, need to be deployed. A considerable amount of German news outlets already uses word filters to prevent potentially illegal comments from being published. However, more reliable methods of automatic detection need to be developed, as the currently used word filters partially filter civil comments.

Future research should further assess the effectiveness of different moderation strategies for different levels and various types of incivility. Does, for example, a discursive moderation of accusations of lying have different effects than the same moderation style in response to insults? Finally, future studies should investigate the effects of incivility and different moderation strategies on other users' issue- and discussion-related attitudes and behavior, for example their willingness to participate, or their perceived responsibility to engage with uncivil comments. In that sense, it is important to find ways how to stimulate users' willingness to complement professional moderation activities by engaging in corrective action themselves (Leonard, Rueß, Obermaier, and Reinemann, this issue).

7.3. Conclusion

Numerous studies have argued that moderation is a crucial element of successful online discussion cultures of almost every kind. Still, some forms of moderation are considered as fairer and more transparent than others. The current study investigated the patterns, determinants, and potential effects of different styles of interactive journalistic moderation. Such moderation could be a transparent and effective strategy to not simply eliminate uncivil statements from the public discourse, but to address them transparently and thereby increase robust civility in online discussions (Garton Ash, 2016). While the current study shows that moderators indeed engaged with commenters who use public-level incivility, the use of personal-level incivility often remained unanswered. Together with the finding that not all deliberative or non-deliberative moderation styles of initial comments were related to less incivility in the subsequent reply comments, the current study provides an important stepping stone for future research on moderation, but at the same time shows that dealing with incivility remains a major and complex challenge for society.

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

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, N.J.: Prentice-Hall.

Bergström, A., & Wadbring, I. (2014). Beneficial yet crappy: Journalists and audiences on obstacles and opportunities in reader comments. *European Journal of Communication, 30*(2), 137–151. https://doi.org/10.1177/0267323114559378

Bodie, G. D., & Burleson, B. R. (2008). Explaining variations in the effects of supportive messages: A dual-process framework. In C. Beck (Ed.), *Communication yearbook, 32* (pp. 355–398). Mahwah, NJ: Lawrence Erlbaum.

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

Chen, G. M., Riedl, M. J., & Huang, Q. E. (2018, November). *Taming incivility in online comment streams*. Paper presented at the 104th annual conference of the National Communication Association, Salt Lake City, US.

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2014, June). *How community feedback shapes user behavior.* Paper presented at the 8. International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI, USA. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8066

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

Diakopoulos, N. A., & Naaman, M. (2011). Towards quality discourse in online news comments. In *CSCW '11 Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 133–142). New York: ACM. https://doi.org/10.1145/1958824.1958844

Eilders, C. (2002). Conflict and consonance in media opinion. *European Journal of Communication, 17*(1), 25–63. https://doi.org/10.1177/0267323102017001606

Esau, K., Friess, D., & Eilders, C. (2017). Design matters! An empirical analysis of online deliberation on different news platforms. *Policy & Internet*, *9*(3), 321–342. https://doi.org/10.1002/poi3.154

euro I topics (2018a). *Berliner Morgenpost.* Retrieved from https://www.eurotopics.net/de/148419/berliner-morgenpost

euro I topics (2018b). *Der Tagesspiegel.* Retrieved from https://www.eurotopics.net/de/148489/der-tagesspiegel

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet, 7*(3), 319–339. https://doi.org/10.1002/poi3.95

Garton Ash, T. (2016). *Free speech: Ten principles for a connected world*. New Haven, London: Yale University Press.

Gastil, J. (2008). *Political communication and deliberation*. Los Angeles, CA: Sage Publications.

Gervais, B. T. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics, 12*(2), 167–185. https://doi.org/10.1080/19331681.2014.997416

Goodman, E., Cherubini, F., & Waldhorn, A. (2013). Online comment moderation: emerging best practices. A guide to promoting robust and civil online conversation. *World Association of Newspapers and News Publishers,* 1–71. Retrieved from http://www.wan-ifra.org/private-download/wan-ifra-online-commenting

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology, 17*(1), 42–109.

Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy* (2nd ed.). Cambridge, MA: Mit Press.

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Philadelphia, PA: Temple University Press.

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S. A. Barab, R. Kling & J. H. Gray (Eds.), *Designing for Virtual Communities in the Service of Learning* (pp. 338–376). New York: Cambridge University Press.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd rev. ed.). New York: Routledge.

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

Hwang, H., Kim, Y., & Kim, Y. (2016). Influence of discussion incivility on deliberation: An examination of the mediating role of moral indignation. *Communication Research, 45*(2), 213–240. https://doi.org/10.1177/0093650215616861

Janssen, D., & Kies, R. (2005). Online forums and deliberative democracy. *Acta Politica, 40*(3), 317–335. https://doi.org/10.1057/palgrave.ap.5500115

Kalch, A., & Naab, T. K. (2017). Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *Studies in Communication and Media (SCM), 6*(4), 397–419.

Kraut, R. E., Resnick, P., & Kiesler, S. (2011). *Building successful online communities: Evidence-based social design.* Cambridge, Mass: Mit Press.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Ksiazek, T. B. (2016). Commenting on the news. *Journalism Studies,* 1–24. https://doi.org/10.1080/1461670X.2016.1209977

Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication Theory, 15*(2), 127–147. https://doi.org/10.1111/j.1468-2885.2005.tb00329.x

Loke, J. (2012). Old turf, new neighbors. *Journalism Practice, 6*(2), 233–249. https://doi.org/10.1080/17512786.2011.616649

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Manosevitch, E. & Walker, D. M. (2009, April). *Reader comments to online opinion journalism: A space of public deliberation*. Paper presented at the 10th International Symposium on Online Journalism, Austin, TX. Retrieved from https://online.journalism.utexas.edu/2009/papers/ManosevitchWalker09.pdf

Massaro, T. M., & Stryker, R. (2012). Freedom of speech, liberal democracy, and emerging evidence on civility and effective democratic engagement. *Arizona Law Review, 54*, 375–441.

Maurer, M., & Reinemann, C. (2006). *Medieninhalte: Eine Einführung*. [Media content: An introduction]. Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90179-4

Mayring, P. (2000). Qualitative content analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 1*(2). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1089/2385

Meedia (2016). Überfordert *vom Leser-Hass: Zeitungsredaktionen schränken Kommentarfunktion ein* [Overwhelmed by readers' hatred: Newspaper editors restrict the comment function]. Retrieved from http://meedia.de/2016/03/01/ueberfordert-vom-leser-hass-zeitungsredaktionen-schraenken-kommentarfunktion-ein/

Meyer, H. K., & Carey, M. C. (2014). In moderation: Examining how journalists' attitudes toward online comments affect the creation of community. *Journalism Practice, 8*, 213–228. https://doi.org/10.1080/17512786.2013.859838

Miron, A. M., & Brehm, J. W. (2006). Reactance theory—40 years later. *Zeitschrift für Sozialpsychologie*, *37*(1), 9–18. https://doi.org/10.1024/0044-3514.37.1.9

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication*, *11*, 3182–3202.

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication, 67*(4), 586–609. https://doi.org/10.1111/jcom.12312

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

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2017). Reuters Institute Digital News Report 2017. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf?utm_source=digitalnewsreport.org&utm_medium=referral

Nielsen, C. E. (2014). Coproduction or cohabitation: Are anonymous online comments on newspaper websites shaping news content? *New Media & Society, 16*(3), 470–487. https://doi.org/10.1177/1461444813487958

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

Pertsch, S. (2016). Social media-charts of German media outlets [Blog post]. Retrieved from https://www.sebastian-pertsch.de/6488/socialmedia-charts-der-medien.html

Prochazka, F., Weber, P., & Schweiger, W. (2018). Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism Studies,* 19 (1), 62–78. https://doi.org/10.1080/1461670X.2016.1161497

Reuter, M. (2016). *Moderation bleibt Handarbeit: Wie große Online-Medien Leserkommentare moderieren* [Moderation remains manual: How big online media moderate reader comments]. netzpolitik.org. Retrieved from https://netzpolitik.org/2016/moderation-bleibt-handarbeit-wie-tageszeitungen-leserkommentare-moderieren/

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, September 2016*. Bochum, Germany.

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior, 58*, 461–470. https://doi.org/10.1016/j.chb.2016.01.022

Rowe, I. (2015a). Civility 2.0: a comparative analysis of incivility in online political discussion. *Information, Communication & Society, 18*(2), 121–138. https://doi.org/10.1080/1369118X.2014.940365

Rowe, I. (2015b). Deliberation 2.0: Comparing the deliberative quality of online news user comments across platforms. *Journal of Broadcasting & Electronic Media, 59*(4), 539–555. https://doi.org/10.1080/08838151.2015.1093482

Ruiz, C., Domingo, D., Micó, J. L., Díaz-Noci, J., Meso, K., & Masip, P. (2011). Public sphere 2.0? The democratic qualities of citizen debates in online newspapers. *The International Journal of Press/Politics, 22,* 463–487. https://doi.org/10.1177/1940161211415849

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice, 8*(1), 18–33. https://doi.org/10.1080/17512786.2013.813194

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century.

Springer, N., Engelmann, I., & Pfaffinger, C. (2015). User comments: motives and inhibitors to write and read. *Information, Communication & Society, 18*(7), 798–815. https://doi.org/10.1080/1369118X.2014.997268

Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. *Journal of Public Deliberation, 3*(1), Article 12.

Stroud, N. J., Scacco, J. M., Muddiman, A., & Curry, A. L. (2015). Changing deliberative norms on news organizations' Facebook sites. *Journal of Computer-Mediated Communication, 20*(2), 188–203. https://doi.org/10.1111/jcc4.12104

Stroud, N. J., van Duyn, E., & Peacock, C. (2016). *News commenters and news comment readers*. Retrieved from http://engagingnewsproject.org/

Stryker, R., Conway, B. A., & Danielson, J. T. (2016). What is political incivility? *Communication Monographs, 83*(4), 535–556. https://doi.org/10.1080/03637751.2016.1201207

Weber Shandwick, Powell Tate, & KRC Research. (2017). Civility in America VII: The state of incivility. Retrieved from http://www.webershandwick.com/uploads/news/files/Civility_in_America_the_State_of_Civility.pdf

Wise, K., Hamman, B., & Thorson, K. (2006). Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate. *Journal of Computer-Mediated Communication, 12,* 24–41. https://doi.org/10.1111/j.1083-6101.2006.00313.x

Wright, S. (2006). Government-run online discussion fora: Moderation, censorship and the shadow of control. *The British Journal of Politics and International Relations, 8*(4), 550–568. https://doi.org/10.1111/j.1467-856X.2006.00247.x

Wüllner, D. (2015). *Lassen Sie uns diskutieren* [Let's discuss]. Retrieved from http://www.sueddeutsche.de/kolumne/ihre-sz-lassen-sie-uns-diskutieren-1.2095271

Ziegele, M. (2016). *Nutzerkommentare als Anschlusskommunikation. Theorie und qualitative Analyse des Diskussionswerts von Online-Nachrichten* [User comments as media-stimulated interpersonal communication. Theory and qualitative analysis of the discussion value of online news]. Wiesbaden: Springer VS.

Ziegele, M., & Jost, P. B. (2016). Not funny? The effects of factual versus sarcastic journalistic responses to uncivil user comments. *Communication Research*, 1–30. https://doi.org/10.1177/0093650216671854

Ziegele, M., Köhler, C., & Weber, M. (2017). *Socially destructive! Effects of hateful user comments on recipients' prosocial behavior.* Paper presented at the 67th Annual Conference of the International Communication Association (ICA), San Diego, USA.

Ziegele, M., Quiring, O., Esau, K., & Friess, D. (2018). Linking news value theory with online deliberation: How news factors and illustration factors in news articles affect the deliberative quality of user discussions in SNS' comment sections. *Communication Research*, 1–31. https://doi.org/10.1177/0093650218797884

Ziegele, M., Weber, M., Quiring, O., & Breiner, T. (2018). The dynamics of online news discussions: Effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions. *Information, Communication & Society, 21*(10), 1419–1435. https://doi.org/10.1080/1369118X.2017.1324505