A matter of perception? investigating subjective and objective exposure to hate speech with a survey and mobile longitudinal linkage study

Wirz, Dominique S.; Blassnig, Sina

DOI
10.1080/1369118X.2025.2461646

Publication date
2025

Document Version
Final published version

Published in
Information Communication and Society

INFORMATION, COMMUNICATION & SOCIETY
2025, VOL. 28, NO. 4, 723–743
https://doi.org/10.1080/1369118X.2025.2461646

Routledge
Taylor & Francis Group

A matter of perception? investigating subjective and objective exposure to hate speech with a survey and mobile longitudinal linkage study

Dominique S. Wirz [a] and Sina Blassnig[b,c]

[a]Amsterdam School of Communication Research, University of Amsterdam, Amsterdam, Netherlands; [b]Department of Communication and Media Research, University of Fribourg, Fribourg, Switzerland; [c]Institute of Digital Communication and Media Innovation, University of Applied Sciences of the Grisons, Chur, Switzerland

ABSTRACT

An increasing number of media users report frequent encounters with hate speech on the internet. Content moderation is only effective when the applied criteria align with users' perceptions of hate speech. To explore what media users perceive as hate speech and which factors influence this perception, we used a multi-methods approach. First, we conducted a survey with a representative sample of the Swiss population (N = 2000). Second, participants who reported frequent exposure to hate speech took part in a two-week mobile longitudinal linkage study, uploading screenshots and answering questions each time they encountered hate speech. We analysed N = 564 screenshots to see if they met common academic definitions of hate speech. Our findings show that impoliteness and insults are more likely to be considered hate speech when they affect one's social identity, and that self-reports indicate higher exposure to hate speech than what was documented via screenshots.

ARTICLE HISTORY
Received 5 August 2024
Accepted 28 January 2025

KEYWORDS
Hate speech; social identity; experience sampling; mobile longitudinal linkage analysis; survey experiment

Hate speech proliferates across various online platforms, including social media, news comment sections, discussion forums, and messaging services. Hate speech encompasses insults, defamation, and threats or incitements to violence targeting specific social groups (e.g., Council of Europe, 2022; Fortuna & Nunes, 2019). Despite efforts to identify and delete hate speech manually or automatically, exposure to it has increased. In a cross-national survey in 2013, 43% of young adults reported exposure to hate speech in the past three months (Hawdon et al., 2017), rising to 71% in 2018 (Reichelmann et al., 2021). In 2021, every second person aged 18-35 in the EU stated that they had been victims of hate speech (HateAid, 2021).

This high exposure to hate speech questions the effectiveness of current detection approaches and if users' perceptions of hate speech align with platforms' or media

CONTACT Dominique S. Wirz d.s.w.wirz@uva.nl.ch Amsterdam School of Communication Research, University of Amsterdam, Postbus 15791, 1001 NG Amsterdam, Netherlands

companies' moderation criteria. Regulations and academic definitions distinguish between hate speech and impoliteness (Friess et al., 2021; Rossini, 2020). Hate speech undermines democratic values by fostering division and discrimination, silencing marginalized voices, and inciting violence and is for example restricted by the European Convention of Human Rights. Impoliteness, in contrast, refers to statements inappropriate in tone that do not qualify as hate speech and thus require other interventions, like counterspeech (Council of Europe, 2022). It is however unclear to what extent media users can distinguish between these two types of expressions, and to what extent they evaluate their severity differently.

Additionally, this distinction may not be decisive for whether users feel personally affected by statements. Media users tend to classify even neutral media reporting as hostile if it refers to a social group they strongly identify with. This bias known as hostile media perception (Vallone et al., 1985) could similarly apply to perceptions of hate speech or impoliteness. Thus, one person might perceive a statement as hate speech while another perceives it as harmless. Since regulatory measures and automated deletion of hate speech neglect perceptional biases, users belonging to certain social groups could be disproportionally exposed to (perceived) hate speech. Moreover, few studies provide representative figures on *how often* citizens encounter hate speech online. Reichelmann et al. (2021) report that between 10% (Poland) and 24% (USA) of young adults frequently encounter hate speech. Yet, it remains unclear what 'frequently' means – is it daily, weekly, or once a month?

From a political communication perspective, understanding users' perceptions of hate speech and impoliteness is crucial. Digital platforms shape public discourse and political participation (e.g., Papacharissi, 2004), so understanding the frequency and nature of citizens' encounters with online hate speech can inform strategies to mitigate its harmful effects. Moreover, understanding which types of statements media users perceive as hate speech can help to improve content moderation, promoting a more inclusive and equitable online environment conducive to democratic principles.

This paper addresses these research gaps by combining survey, data donation, and content analysis in Switzerland.[1] First, a representative survey in German- and French-speaking Switzerland (N = 2000) investigated how often participants subjectively encounter hate speech. To augment self-reports, participants were shown specific examples, for which they indicated how frequently they encounter such content types. Second, we conducted an online experience sampling study involving 119 participants from the survey adopting a mobile longitudinal linkage approach (Otto et al., 2022). Over two weeks, participants documented hate speech encounters via screenshots. The donated screenshots (n = 567) were examined in a manual quantitative content analysis to assess hate speech and impoliteness. Both studies were pre-registered prior to collecting the data[2] and approved by the IRB of the University of Fribourg [Application Nr. SES IRB 2023-02-01]. They advance our understanding of media users' perceptions of hate speech by a) measuring to what extent media users evaluate different statements as 'hate speech', b) testing how social identities influence this evaluation, and c) comparing self-reported exposure frequency to hate speech (subjective exposure) with objective exposure frequency measured with the mobile longitudinal linkage approach.

Literature review

User perceptions of hate speech

Hate speech is a collective term for various forms of public expressions of hatred or devaluing attitudes, such as insults or disparagement based on social group membership (Calvert, 1997; Rossini, 2020; Siegel, 2020). Fortuna and Nunes (2019) define hate speech based on four characteristics: 1) it refers to specific individuals or groups, 2) intends to incite violence or hatred, 3) attacks or denigrates, and 4) can be subtle or humorous. Sellars (2016) adds that hate speech causes harm, which is intended by the sender of the message. The Council of Europe (2022) distinguishes between insults, defamation, and incitement to violence or threats. In Switzerland, hate speech is punishable under the Swiss Criminal Code (SCC) as 'insult' (SCC Art. 177), 'defamation' or 'slander' (SCC Art. 173-174), 'public incitement to crime or violence' (SCC Art. 259), or 'discrimination and incitement to hatred' (SCC Art. 261). Thus, we define hate speech as insults, defamation, or threats and incitement to violence targeted at collectives or individuals based on group-defining characteristics (e.g., race, gender, religion) rather than individual attributes.

Although hate speech can be subtle or humorous, it must be distinguished from statements that are impolite in tone but do not fulfill the characteristics of hate speech, such as disdain or condescension, sarcasm, 'shouting' (use of capital letters or exclamation points), vulgar language, name-calling, or attacks on arguments (Friess et al., 2021; Papacharissi, 2004; Rossini, 2020).[3] Impolite statements are not necessarily hate speech, and hate speech is not always impolite. Since impolite statements violate norms of civil discussions, they are sometimes referred to as incivility (e.g., Friess et al., 2021; Kenski et al., 2020; Muddiman, 2017). However, incivility goes beyond mere impoliteness and includes statements qualifying as hate speech (Stoll et al., 2020). Therefore, we use the terms hate speech and impoliteness to differentiate between harmful statements that are potentially restricted by law (hate speech), and those inappropriate in tone but requiring alternative responses, such as counter-speech (impoliteness).

Previous research indicates many internet users frequently encounter hateful content online (Hawdon et al., 2017; Reichelmann et al., 2021), but it remains vague how often 'frequently' exactly is. Using a more differentiated scale, a representative study of Swiss adolescents aged 12 to 19 shows that 48% of respondents encounter hate speech several times a week and 12% several times a day (Külling et al., 2021). Hate speech thus appears to be widespread in young people's media repertoires in Switzerland. However, no representative data exists for the general adult population in Switzerland, and also international surveys have focused on youths and young adults. Since social media use varies by age (Reuters, 2023), it is important to widen the focus regarding perceptions of hate speech.

Furthermore, few studies have investigated how often users perceive different forms of hate speech and impoliteness online. The difference between hate speech and impoliteness is crucial theoretically and legally, but may be difficult for users to recognize. We do not want to say that the definition of hate speech is in the eye of the beholder, as this harbors the danger or relativizing hate speech. There are objective theoretical and legal definitions. However, user perception of hate speech can be influenced by message, user, or context characteristics (Kenski et al., 2020; Muddiman, 2017; Schmid et al.,

2022). In terms of message characteristics, severe forms, e.g., threats of violence, are more likely to be evaluated as hate speech (Schmid et al., 2022). Among forms of impoliteness, name-calling was found to be perceived the most severe (Kenski et al., 2020) and might thus be considered as hate speech by some. Regarding user characteristics, females and persons scoring high on the personality trait agreeableness tend to evaluate impoliteness as more severe, while politically conservatives perceive it as less severe (Kenski et al., 2020); this likely also applies to hate speech. Further, the more time users spend on social media, the less they perceive offensive statements as hate speech (Schmid et al., 2022). Context also matters; statements by politicians from a party one identifies with are perceived as more civil than those of politicians from other parties (Muddiman, 2017). Moreover, repeated exposure to hate speech leads to desensitization; hate speech is evaluated as less severe, and prejudice against members of the outgroup increases (Soral et al., 2018). Thus, frequently encountered forms of hate speech may be perceived as more acceptable. Overall, various factors influence whether an offensive statement is perceived as hate speech.

Hate speech and social identity

In addition to the challenge of distinguishing hate speech from impoliteness, our perceptions of media content are biased. Particularly relevant here is the hostile media effect, which refers to perceiving neutral media reporting as hostile towards one's own social group, a phenomenon especially observed for minorities (Vallone et al., 1985).

Hostile media perceptions are closely related to social identity theory (Kim & Hwang, 2019), which suggests people categorize themselves and others into distinct social groups – in-groups and out-groups, with a positive bias towards the in-group. Hostile media perceptions reflect individuals' in-group bias or out-group hostility when evaluating media content or statements relevant to their social identity (Ariyanto et al., 2007). Individuals strive to maintain a positive image of their in-group by selectively recalling and categorizing media content. Furthermore, users may apply different standards when evaluating the strength or weakness of arguments, focusing on aspects favorable to their in-group and disregarding unfavorable ones.

These biases can lead individuals to view neutral or even positive statements as hostile, especially when directed towards a group they strongly identify with (Vallone et al., 1985). This phenomenon is amplified when content makes people's social identity salient (Reid, 2012). In the context of social media, attacks on a group have been shown to make users' social identity salient, promoting hostile media perceptions (Cohen et al., 2020; Kim & Hwang, 2019). Thus, people may be more likely to perceive statements as hate speech if they identify with the derogated group. For example, Papcunová et al. (2023) find that people who feel closer to migrants more likely perceive statements targeting migrants as hate speech. While Wojatzki et al. (2018) find no significant gender difference in perceptions of hate speech towards women, Pedersen et al. (2023) find that both female and male politicians evaluate hate speech targeted at women more severe than hate speech targeted at men. Thus, akin to hostile media perceptions in traditional contexts (Hartmann & Tanis, 2013), the targeted group's status seems to influence hate speech perceptions; if the status of the group one identifies with is perceived as lower than the out-group's, this may foster hostility perceptions. However, since hate speech attacks

a group's status, we assume that majority groups also more likely perceive statements targeted at their social identity as hate speech.

Building on social identity theory and hostile media perceptions, research further suggests that belonging to the targeted social group of hate speech may influence support for punitive measures against the perpetrator. Hostile media perceptions are fostered by cognitive and affective involvement (Matthes, 2013). Perceiving media content as biased against one's in-group can evoke negative emotions like anger, triggering specific behavioral tendencies (Arpan & Nabi, 2011). Anger has been linked to a preference for punitive responses (Kühne & Schemer, 2015; Nabi, 2003), suggesting individuals belonging to the target group of hateful or impolite statements may be more supportive of punishing the sender due to the anger evoked by the perceived offense towards one's in-group.

Research design

Based on these theoretical considerations, we investigate the perception of hate speech and impoliteness in the Swiss population. First, we conducted a representative survey to measure the frequency of exposure to hate speech in general, and to different forms of hate speech and impoliteness more specifically. Based on the survey responses, we identified individuals regularly exposed to hate speech and invited them to participate in a mobile experience sampling study. In this second study, participants documented all instances of hate speech they encountered over two weeks. We analyzed the collected screenshots using content analysis to categorize them as hate speech or impoliteness. In combination, the two studies allow us to compare different approaches of measuring exposure to hate speech and gain a better understanding of what media users perceive as hate speech. The two studies and their research questions are described in the following sections.

Study 1: Representative survey

The first study comprised a representative survey of the Swiss population (N = 2,000, 18-79 years, internet users, German-speaking and French-speaking Switzerland) to determine the extent of hate speech confrontation. Similar to previous studies (e.g., Külling et al., 2021), respondents were first asked directly how often they perceive hate speech on the internet. They were then asked which groups they perceive to be targeted by hate speech, how often they themselves feel affected, and what counter measures they consider appropriate. As it remains unclear from these self-reports what exactly the respondents mean by hate speech, they were additionally shown specific examples.

First, the survey serves descriptive purposes, addressing the following research questions:

RQ1: How often and where do Swiss media users encounter hate speech?

RQ2: To what extent do Swiss media users encounter different types of hate speech (insult, defamation, call for violence) and impoliteness?

RQ3: To what extent do Swiss media users consider different types of messages (insult, defamation, call for violence) and impoliteness as 'hate speech'?

RQ4: How do Swiss media users evaluate different types of hate speech (insult, defamation, call for violence) and impoliteness?

Second, we included an experimental component in the survey to test if social identities affect hate speech perceptions and evaluations. For the examples shown to participants, we varied the derogated groups, allowing us to test if participants are more likely to categorize statements as hate speech when they belong to the derogated group. Based on the hostile media effect, we propose the following two hypotheses:

H1: When individuals belong to the target group of an offensive statement (hate speech or impoliteness), they are more likely to perceive the statement as hate speech.

H2: When individuals belong to the target of group of an offensive statement (hate speech or impoliteness), they are more supportive of punitive measures against the sender.

Sample

Recruitment for the survey was carried out by the market research institute Intervista[4] from May 2 to 11, 2023, involving 2000 participants from German- and French-speaking Switzerland. The sample is representative regarding age, gender, and education for these language regions, with weighting factors applied in the following *descriptive* analyses to compensate for minor deviations from quotas. Table 1 shows both unweighted and weighted demographic characteristics of respondents. On average, the survey took approximately 10 min to complete (M = 11 min, Mdn = 9 min).

Operationalization

Participants received the following introduction and definition of hate speech:

> Sometimes people are hostile to other people online because they belong to a certain population group, for example due to skin colour, language, nationality, religion, gender, sexual orientation, physical or mental disability, physical appearance or education, income and profession. These people can be insulted or threatened, for example. Such statements are referred to as hate comments in the following. This includes comments that are unpleasant but permissible in terms of freedom of expression, as well as potentially criminal statements.

Exposure to hate speech in general, was measured by asking participants how often they personally encounter hate speech on the Internet on a scale from 1 = 'never' to 8 = 'several times a day'.

Types of platforms: We further asked about exposure to hate speech for the following types of platforms: social media (e.g., Facebook, Twitter, Instagram, TikTok, LinkedIn, etc.), video platforms (e.g., YouTube), news websites (e.g., SRF News, Blick, 20 Minuten), discussion forums (e.g., Reddit, Discord, Quora), messenger apps (e.g., WhatsApp, Telegram, Signal), private messaging channels (e.g., email, SMS), shopping websites (e.g., Amazon, Etsy).

Examples of incivility, insults, defamation, and threats: To measure exposure to and evaluation of different types of hate speech we showed participants examples of mere impoliteness, insults, defamation, and threats. We modified statements containing hate speech or impoliteness found on the internet to correspond to only one of these categories (see Table 2). We created two examples per category. Participants in the survey were randomly shown one example of each category.

Social identity manipulation: To prevent bias in the evaluation of the examples and test the social identity hypothesis, we randomly varied the attacked target groups. For

Table 1. Composition of the sample.

		unweighted		weighted	
		N	%	N	%
Age group	18-29	375	18.5	364	18.0
	30-44	564	27.8	568	28.0
	45-59	572	28.2	571	28.0
	60-79	515	25.4	522	26.0
Gender	diverse	11	0.5	11	0.5
	male	992	49.0	1009	50.0
	female	1023	50.5	1007	50.0
Highest educational qualification (ISCD 2011)	Primary or lower secondary education (ISCED 1 + 2)	86	4.2	85	4.2
	Higher secondary and post-secondary non-tertiary education (ISCED 3-5)	830	38.1	826	40.8
	Tertiary education (ISCED 6-8)	1110	54.8	1115	55.0
Type of settlement	Rural	739	36.5	737	36.0
	Urban	1287	63.5	1289	64.0
Language region	German-speaking	1501	74.1	1506	74.0
	French-speaking	525	25.9	520	26.0
Sexual orientation	heterosexual	1785	88.1	1785	88.0
	homosexual	78	3.8	78	3.8
	bisexual	70	3.5	70	3.5
	asexual	11	0.5	11	0.5
	other	12	0.6	12	0.6
	don't know / no answer	70	3.5	70	3.5
Citizenship	exclusively Swiss since birth	1489	73.5	1491	74.0
	Swiss since birth with double citizenship	175	8.6	174	8.6
	naturalized Swiss	208	10.3	208	10.0
	other citizenship	154	7.6	153	7.6
Religion	non-denominational	744	36.7	745	37.0
	catholic	591	29.2	591	29.0
	protestant	527	26.0	527	26.0
	orthodox	13	0.6	13	0.6
	jewish	6	0.3	6	0.3
	muslim	25	1.2	25	1.2
	buddhist	9	0.4	9	0.4
	hinduist	5	0.2	5	0.2
	other	67	3.3	67	3.3
	don't know / no answer	39	1.9	39	1.9
Political orientation	left	470	23.2	469	22.9
	center	1108	57.3	1162	57.5
	right	237	11.7	240	11.8
	don't know / no answer	157	7.7	155	7.6
Disabilities	no restrictions	1748	86.3	1748	86.0
	some restrictions	242	11.9	242	12.0
	severe restrictions	36	1.8	36	1.8

impoliteness, insults, and threats, the target groups included nationality (Swiss vs. foreigners), political affiliation (left vs. right), religion (Christian vs. Muslim), and sexual orientation (hetero- vs. homosexual). For defamation, we used more specific groups, i.e., political (the leadership of the social democratic party (SP) vs. the leadership of the Swiss people's party (SVP)) or religious (the catholic bishops vs. the Islamic central council) entities, as defamation requires verifiable claims. Thus, each of the four examples of hate speech or impoliteness presented in the survey, randomly featured attacks on majority or minority groups.

Perception and evaluation of hate speech: For each of the example statements, participants were asked to indicate their agreement to the following items on a 5-point Likert scale from 'do not agree at all' to 'fully agree': *I perceive this statement as hate speech; I*

Table 2. Examples of hate speech and impoliteness shown in the survey.

	Statement	Target groups
Insult	I demand a ban on these [target group] idiots and the cessation of all payments from the taxpayer's purse to such or similar 'associations'!	leftist, rightist, gay, heterosexual, Muslim, Christian, foreign, Swiss
	The [target group] are mostly still dumb as rocks and otherwise completely out of place in a modern merit-based society.	leftists, rightists, gays, heterosexuals, Muslims, Christians, foreigners, Swiss
Defamation	Why should Switzerland pay for this pedophile association? The [target group] have already amassed enough wealth!	The leadership of the SP / SVP, the catholic bishops, the Islamic Central Council
	Embarrassing, corrupt clowns, these [target group].	The leaders of the SP / SVP, catholic bishops, members of the Islamic Central Council
Threat	Fat filthy pigs, I can smell them from here. Hopefully, these [target group] will get their faces smashed in self-defense one day.	leftists, rightists, gays, heterosexuals, Muslims, Christians, foreigners, Swiss
	As punishment for their behavior, these [target group] belong in the gas chamber.	leftists, rightists, gays, heterosexuals, Muslims, Christians, foreigners, Swiss
Impoliteness	Here's my feedback: The opinion of these [target group] is nothing more than intellectual diarrhea.	leftists, rightists, gays, heterosexuals, Muslims, Christians, foreigners, Swiss
	It's probably just a 'fart' from these [target group]. Disgusting, disgusting, shame, shame.	leftists, rightists, gays, heterosexuals, Muslims, Christians, foreigners, Swiss

often see such statements when I use media; I think such statements should be deleted; I think the author of such statements should be reported to the police.

Participant's social identity: To test the social identity hypothesis, it was necessary to know if participants belong to a group targeted in the statement they saw. We therefore measured participants' *sexual orientation*, *nationality*, *religion*, and *political orientation*, then created dichotomous variables for each target group (belongs to target group vs. not). For sexual orientation, participants identifying neither as homo- nor hetero-sexual were excluded; for nationality, Swiss participants with another nationality were considered as foreigners; for religion, participants with religions other than Christian or Muslim (or no religion) were not included in either group; and for political orientation, participants scoring 1-3 on a nine-point scale were considered as left-wing, and those scoring 9-11 as right-wing.[5]

Findings

Regarding *how often* Swiss citizens encounter hate speech (*RQ1*), the representative survey shows that 69% of Swiss internet users have already seen hate speech online. Whereas 31% stated that they had never seen hate speech, around a third of the population frequently encounter it: 1.9% several times a day, 6.3% daily, 16% several times a week and 10% once a week. Another third experiences hate speech occasionally: 7% once a fortnight, 9.8% once a month and 17% less than once a month.

Regarding the second part of *RQ1* (*where*), Figure 1 shows that hate speech is most frequently encountered on social media, followed by news websites and video platforms. Swiss people perceive hate speech less frequently on shopping websites, discussion forums, and via private messaging channels like email or SMS. On the one hand, this may be because these channels are used less frequently overall (and if one only visits a channel once a month, it is impossible to encounter hate speech on that channel more often). On the other hand, it may also be related to the nature of content or the way

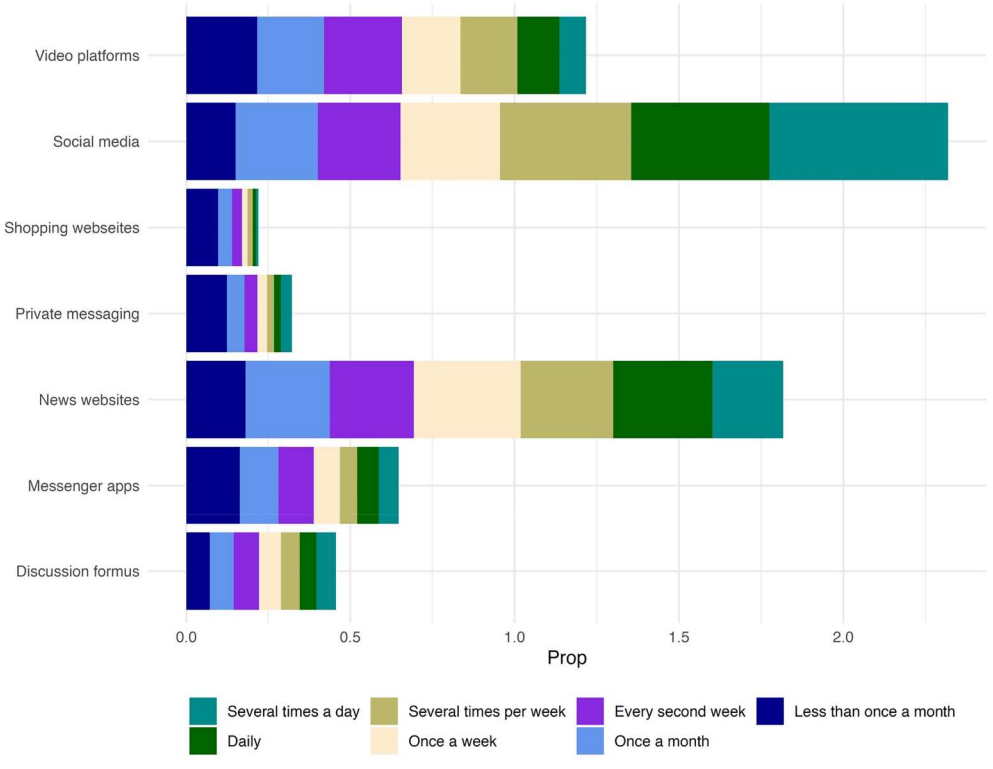

Figure 1. Exposure to hate speech by platform.

these channels are used. Users perceiving hate speech daily primarily encounter it on social media.

To investigate *RQ2* on the relative exposure to different types of hate speech (insult, defamation, call for violence) and impoliteness, respondents who had perceived hate speech before were shown four examples: insulting, defamatory, threatening, and impolite statements. For each statement, respondents indicated whether they frequently see such statements online. Figure 2 shows that the respondents most commonly perceive insults, with 43% fully or somewhat agreeing they frequently perceive such statements. Regarding defamation, 39.8% fully or somewhat agree to see them often, while 29% report this for threats, and 38.7% for impolite statements.

To address *RQ3*, respondents were asked to indicate the extent to which they perceive the examples as hate speech. Looking at the type of statements independently of the attacked groups (Figure 3) shows that the Swiss population differentiates between forms of hate speech. For insults, 76% fully or somewhat agree it qualifies as hate speech, compared to 59% for defamation and 94% for threats. For impolite statements, 50% (somewhat) agree it constitutes hate speech. Thus, the population can partly differentiate between hate speech and impoliteness, although defamatory statements are not perceived as different from impoliteness as insults and threats.

Perceptions of statements as 'hate speech' are also reflected in the evaluations of the statements (*RQ4*), specifically regarding the desired consequences. Respondents were asked for each example statement whether it should be deleted and if the author should

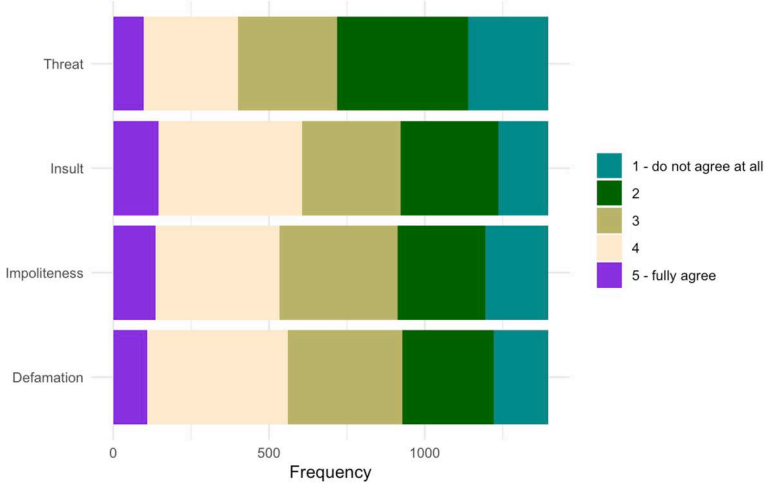

Figure 2. Exposure to different types of hate speech and impoliteness (agreement to the item: *I often see such statements when I use media*).

be reported. For threats, 86% of respondents fully or somewhat agree that such statements should be deleted, compared to 63% for insults. For defamation (46%) and impoliteness (44%), less than half of respondents agree with deletion. Additionally, 66% fully or somewhat agree that authors of threats should be reported to the police, while a minority hold this opinion for insults (28%) and defamation (22.1%). Approval to report impolite statements to the police (21.3%) is almost as high as for defamation.

To test *H1* if participants are more likely to perceive statements as hate speech when the statement attacks a group participants identify with, we compared the mean perception of statements as hate speech between participants who saw a version of the statement attacking a social group they belonged to or and those who did not. Table 3 shows these

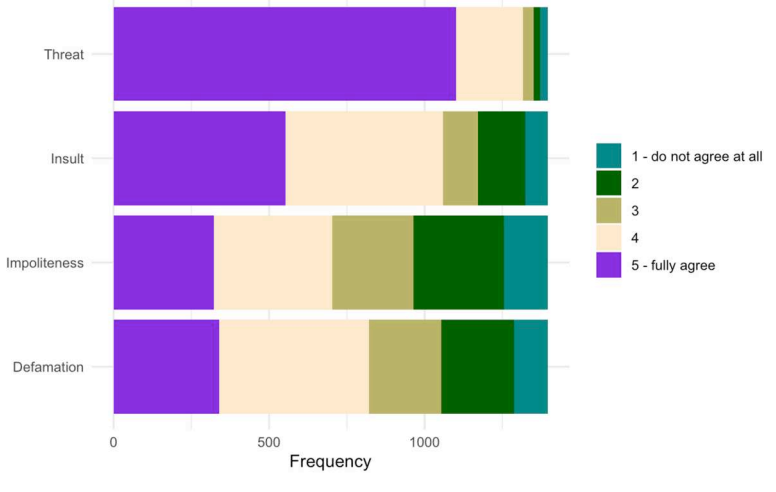

Figure 3. Evaluation of statements as hate speech (agreement to the item: *I perceive this statement as hate speech*).

Table 3. Perceptions of statements as hate speech by group membership.

	Social group membership (social identity)			
	targeted		not targeted	
	n	*M (SD)*	*n*	*M (SD)*
Insult	396	4.19 (1.03)	1002	3.85 (1.21)
Defamation	125	3.66 (1.24)	1273	3.49 (1.24)
Threats/violence	396	4.68 (0.76)	1002	4.68 (0.75)
Impoliteness	384	3.56 (1.21)	1014	3.24 (1.33)

Table 4. Support of punitive measures by group membership.

	Social group membership (social identity)			
	targeted		not targeted	
	n	*M (SD)*	*n*	*M (SD)*
Insult	396	2.85 (1.29)	1002	2.68 (1.29)
Defamation	125	2.61 (1.27)	1273	2.50 (1.23)
Threats/violence	396	3.77 (1.25)	1002	3.84 (1.18)
Impoliteness	384	2.70 (1.26)	1014	2.39 (1.28)

comparisons by the statement type. ANOVAs[6] were conducted for each statement type individually. While threats ($p = 0.878$) and defamation ($p = 0.143$) are evaluated similarly by targeted and not targeted participants, the evaluation of insults ($p < 0.001$, $\eta^2 = 0.02$) and impoliteness ($p < 0.001$, $\eta^2 = 0.01$) is significantly affected by one's social identity. Participants belonging to the attacked group were more inclined to agree that these statements qualify as hate speech, but effect sizes are very small. We thus find mixed evidence for H1: Social identity influences the perception of insults and impoliteness, but not of threats and defamation.

To test *H2,* which states that participants are more supportive of punitive measures when statements target their social identity, we conducted the same analysis with the item '*the author of this statement should be reported to the police*' as the dependent variable. Table 4 shows these comparisons by statement type. The ANOVAs demonstrate that social identity affects the support of punitive measures for insults ($p = 0.02$, $\eta^2 = 0.003$) and impoliteness ($p < 0.001$, $\eta^2 = 0.01$), but not for threats ($p = 0.33$) and defamation ($p = 0.34$).[7] The evidence for H2 is thus mixed, following a similar pattern as for H1.

Discussion

When confronted with example statements, participants clearly identify threats/violence and insults as hate speech, while defamations and especially impoliteness are more contested. In line, participants only support punitive measures for threats, suggesting a lack of awareness in the Swiss population that insults and defamation can also be criminal offenses. Our analysis further demonstrates that individuals tend to perceive insults and impoliteness as more hostile when they belong to the attacked group, unlike defamation and threats. For threats, which are unequivocally considered to be hate speech, it seems intuitive that participants make no distinction as to who is being attacked. For more ambiguously evaluated forms of incivility, group membership seems to bias the perception, except for defamations, possibly due to our operationalization of defamation.

Since defamation can only be legally prosecuted when statements are falsifiable, we chose specific actors as targets of these statements. This approach has likely reduced identification with the mentioned group; for instance, even participants identifying as Catholics may not necessarily identify with catholic bishops. Future research should test if other operationalizations lead to different outcomes. Generally, the selection of examples might have affected our results. Study 2 addresses this shortcoming by using a data donation approach to examine real examples of hate speech in participants' media diets.

Study 1 has additional limitations. First, providing a short definition of hate speech before asking about general frequency of exposure was necessary, but may have affected the evaluation of the example statements. However, given the considerable variation in evaluations of the examples, the potential influence seems to be minimal. Second, self-reported frequency of exposure is likely biased, as participants may find it difficult to estimate. Study 2 addresses this limitation by using a different approach to measure exposure.

Study 2: Mobile experience sampling study

To further explore the extent and form of hate speech Swiss people encounter in their everyday lives, we conducted an experience sampling study. Over 14 days, participants documented encounters with perceived hate speech. This type of online diary study, following the mobile longitudinal linkage analysis approach (Otto et al., 2022), offers several advantages: It does not rely on participants' memory (unlike traditional surveys as our study 1), captures content perceived as hate speech from a user perspective across different platforms (unlike traditional content analysis, where researchers select the analyzed media content), and includes content viewed on mobile devices, apps, and social media (unlike web tracking, which is usually bound to one device).

From the sample in study 1, we recruited 150 people who reported encountering online hate speech several times a week or more. Unlike study 1, study 2 focused on German-speaking participants and aimed for a good mix in terms of gender and age rather than representativeness.[8] Data collection started between May 2 and 22, 2023, and lasted 14 days from the individual start date.

Following an event-driven approach (Otto et al., 2022), participants were asked to take and upload a screenshot each time they perceived hate speech in their daily internet use during the two-week study period, regardless of the platform or source. Post-upload, participants completed a short questionnaire about the statement's source, platform, personal affectedness, and perceived severity (see questionnaire on OSF). Daily reminders were sent, and participants were asked to complete the study even if no hate speech was encountered (i.e., no screenshots uploaded) during this period. With this approach, we can address the following research question:

> *RQ5:* To what extent do messages that users encounter in their daily media use and consider as hate speech correspond to academic and legal definitions of hate speech or impoliteness?

Sample

Of the 150 participants who consented to participate, 119 completed the experience sampling study (Table 5). This sample, consisting of people frequently confronted

with hate speech online according to study 1, is not representative of the Swiss population. Participants in study 2 were younger ($M = 34$, $SD = 11.2$, $min = 18$, $max = 61$) and slightly more male (56.3%, $n = 67$) than the general population (see Table 2). They also used various media channels, especially video platforms, social media, messenger apps, and forums, more frequently than the average population.

Content analysis: operationalization and reliability

The 119 participants who completed the experience sampling study uploaded a total of 575 screenshots, n = 564 without duplications. This media content was then analyzed by the authors in a manual quantitative content analysis regarding the occurring social groups, the content, and the form of hate speech or impoliteness.

Hate Speech: Based on previous research (Stahel, 2020), reports by the Council of Europe (2022) and the Swiss Criminal Code (SCC)[9], we distinguished between a) *insult* (abuse or disparagement), b) *defamation* (slander or libel), and c) *threats or incitement to violence*. Each category was assigned indicators from previous hate speech content analyses (Friess et al., 2021; Rossini, 2020; Ziegele, 2016).

Insults were measured using nine indicators: 1) *Insults related to group identity:* harassment, insults, humiliation, or attacks on people based on their group affiliation. 2) *Personal attacks:* attacks based on personal characteristics, traits, choices, etc. 3) *Dehumanization:* dehumanizing verbal aggression, objectification, or derogatory (e.g., animal-related) names. 4) *Political intolerance:* discrediting opposing political views or actors as invalid or illegitimate, denying their right to exist, or accusing them of an extremely negative societal impact. 5) *Racism:* discrimination, stereotyping, or hateful language based on race or origin. 6) *Socio-economic intolerance:* discrimination, stereotyping, or hateful language based on education, social status, or income. 7) *Sexism or intolerance of sexual freedom:* discrimination, stereotyping, or hateful language based on gender, sexual orientation, or sexual identity. 8) *Religious intolerance:* discrimination, stereotyping, or hateful language based on religious beliefs. 9) *Threats to individual rights:* Denial of equal rights to individuals or groups.

Defamation was differentiated into three indicators: 1) *Lies or falsehoods:* False or distorted claims about individuals or groups based on their group membership. 2) *Institutional defamation:* False or distorted claims, or pejorative language towards policy,

Table 5. Composition of the experience sampling study sample.

		n	%
Age group	18-29	50	42.0
	30-44	48	40.3
	45-59	19	16.0
	60-79	2	1.7
Gender	male	67	56.3
	female	52	43.7
Highest educational qualification (ISCD 2011)	Primary or lower secondary education (ISCED 1 + 2)	4	3.4
	Higher secondary and post-secondary non-tertiary education (ISCED 3-5)	52	43.6
	Tertiary education (ISCED 6-8)	62	52.1
Type of settlement	Rural	40	33.6
	Urban	79	66.4

institutions, or organizations to deprecate or question their credibility or legitimacy. 3) *Offensive stereotypes:* Simplified negative clichés of a person or group.

Threats or incitements to violence were coded as one category: announcements of negative or aggressive sanctions or calls to violent acts or crimes against a recipient.

Impoliteness was coded using six indicators based on Friess et al. (2021), Rossini (2020), and Ziegele (2016): 1) *Disdain/condescension:* negative remarks about people, groups, or ideas that do not attack anyone's honor. 2) *Sarcasm/cynicism*: cynical remarks, mockery, or derogatory humor. 3) *Shouting:* use of capital letters or exclamation points. 4) *Profane or vulgar language*: obscene, foul, or rude language that is inappropriate for professional discourse. 5) *Derogatory names*: offensive, deprecating labels for individuals, groups, or institutions not classified as hate speech. 6) *Attacks on arguments:* Statements attacking, disqualifying, or dismissing a viewpoint or argument.

Social groups (social identity) were coded based on characteristics mentioned in the uploaded screenshots (Bornschier et al., 2021; Wirth et al., 2019), distinguishing: 1) physical appearance, 2) gender identity, 3) sexual orientation, 4) nationality or origin, 5) ethnicity or race, 6) physical or mental disability or illness, 7) political views, 8) religion or belief, 9) language, 10) education, 11) income or occupation.

All indicators were coded as present (1) or absent (0) in the screenshots. Multiple indicators could be coded per screenshot, but each part of a screenshot (words, sentences) could only be assigned to one category. The full codebook is on OSF. Reliability was ensured through several pretests and an intercoder reliability test with 30 screenshots. All categories achieved satisfactory reliability, with Brennan & Prediger's $Kappa < 0.6$ except *sarcasm/cynicism* and *derogatory names* (both $Kappa = 0.53$).

Findings

Study 2 mainly addressed *RQ5,* examining to what degree user-identified hate speech in their daily media use aligns with academic and legal definitions. Our content analysis shows that 48% (n = 271) of the uploaded screenshots contained hate speech as per our codebook, while 66.8% contained impoliteness (n = 377). Both hate speech and impoliteness were present in 26.6% (n = 150), 40.2% (n = 227) were impolite without hate speech, and 11.7% (n = 66) contained neither.

Study 2 provides additional observational insights regarding how often and where Swiss citizens encounter hate speech (*RQ1*). On average, participants uploaded almost five screenshots ($M = 4.8$, $SD = 6.0$, $Median = 3$). Most uploaded only a few, while individuals contributed a large number of uploads, with the maximum being 30 uploads by one person. Most screenshots were taken on social media ($n = 394$, 68.5%) followed by news websites ($n = 101$, 17.9%), and messenger apps ($n = 37$, 6.4%).

Table 6 compares the number of participants who, according to the survey, are confronted with hate speech online several times a week, daily, and several times a day with the respective objective confrontation based on the number of examples they uploaded in the experience sampling study. Based on the number of uploads (≥ 3), over half of the participants (53.8%, $n = 64$) seem to encounter hate speech at least several times a week. However, considering only the uploads coded as hate speech, less than a third (27.7%, $n = 33$) actually encountered hate speech that often in these two weeks (Table 3).

Table 6. Comparison of the confrontation with hate speech (nr. of participants per cell).

	Nr. of uploads					
According to the survey	0 uploads	1-2 uploads	3-12 uploads	13-15 uploads	16 or more	Total
Several times a week	17	22	32	2	6	79, 66.4%
Daily	4	8	12	2	2	28, 23.5%
Several times a day	2	2	5	1	2	12, 10.1%
Total	23	32	49	5	10	119
	19.3%	26.9%	41.2%	4.2%	8.4%	100.0%

Note: Cells highlighted in grey indicate a match between self-reports and nr. of uploads, cells to the left of those indicate an overestimation, cells to the right an underestimation in the survey.

Of the 79 participants reporting encounters with online hate speech several times a week in the survey, 32 uploaded three to twelve examples, supporting their self-reports. However, considering the number of uploads coded as hate speech, only 18 encountered hate speech several times a week. Of the 28 participants reporting daily encounters in the survey, four uploaded enough screenshots ($\geq$ 13) to support their self-reports, but none uploaded that many screenshots coded as hate speech. Of the twelve participants reporting multiple encounters per day in the survey, only two uploaded 13 or more examples coded as hate speech. Around a fifth (n = 23) uploaded no screenshots, suggesting no encounters with hate speech during the study. While this may suggest that participants overestimate their exposure to hate speech in self-report measures, the uploads may also underestimate the exposure due to participants' forgetfulness, laziness, or reluctance to share some instances.

Regarding different types of hate speech (*RQ2)*, insults are most common (41.3%, n = 233), followed by defamation (13.5%, n = 76). Threats or calls to violence are the rarest top category (6.9%, n = 39). Figure 4 shows the frequencies of the subcategories of hate speech and impoliteness.

Post-upload, participants rated the severity of statements from rather harmless (1) to very bad (10) (M = 5.8, SD = 2.2, Median = 6). To explore whether these evaluations differ for screenshots coded as hate speech and directed against one's own group, we calculated a linear regression model predicting the perceived severity with hate speech and personal involvement as independent variables. Statements coded as hate speech (b = 1.01, SD = 0.18, $p < .001$) and statements against one's own group (b = 1.00, SD = 0.02, $p < .001$) are perceived as significantly more severe (F = 1535, $p < .001$, R2 = 0.84, n = 564).

Discussion

The findings from study 2 show a discrepancy between self-reported and self-documented frequency of hate speech exposure: while all participants had reported seeing hate speech several times a week in study 1, only a third uploaded enough screenshots to support this in study 2. Considering both the overall number of uploads and those coded as hate speech, many participants seem to have overestimated their online exposure to hate speech in study 1. Several participants mentioned in their open comments at the end of study 2 that they had seen less hate speech than they had expected, supporting the interpretation that the survey measure overestimated exposure. However, the experience sampling method might have also underestimated exposure because participants forgot to take or upload some screenshots or did not want to share everything they saw (Otto

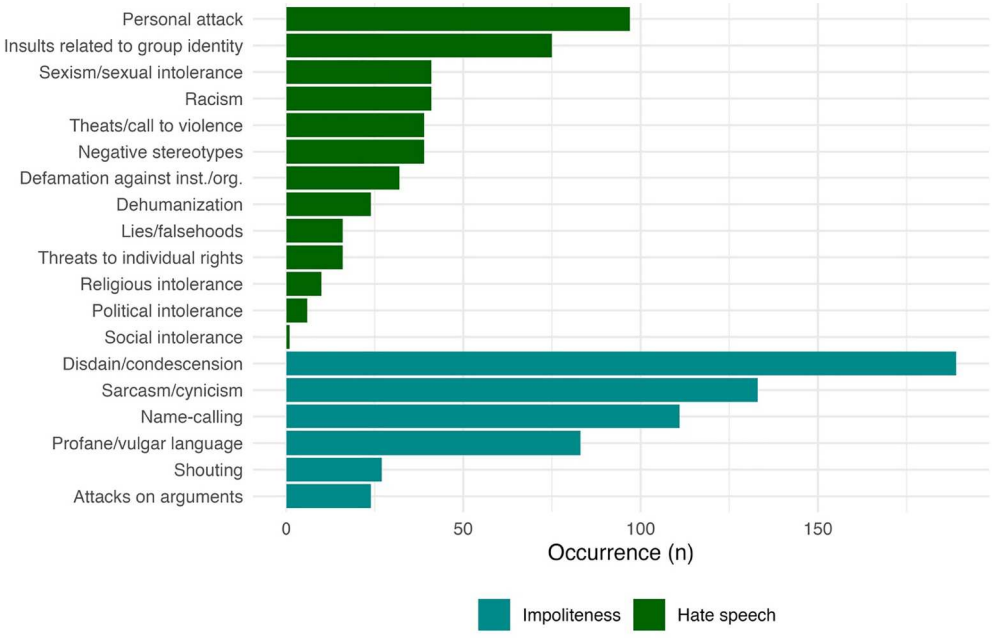

Figure 4. Occurrence of hate speech and impoliteness subcategories in the uploaded screenshots.
Note: Screenshots can contain several categories.

et al., 2022). The true exposure likely lies somewhere in between, requiring more research to determine the validity and reliability of different approaches to measure exposure to hate speech.

Another key finding of study 2 is that more uploaded statements were coded as impoliteness, not hate speech, suggesting that participants often interpreted impoliteness as hate speech. Echoing findings from study 1, insults were most common, followed by defamation, with threats or calls for violence being less frequent. Furthermore, study 2 affirms that hate speech is perceived as more severe than impoliteness, and statements against one's own group are perceived as more severe than those against other groups.

Study 2 has further limitations. Besides the possible non-upload of relevant screenshots, it was not always clear which aspects of the screenshots were seen as hate speech. Similarly, a lack of context may have made some discriminatory statements appear innocuous – which may especially apply to cases neither coded as hate speech nor impoliteness. Finally, the non-representative sample and the high demands of the study design on participants (Otto et al., 2022) limit the generalizability of the findings.

General discussion

Previous research has suggested an increase in exposure to hate speech, despite platforms' and media companies' efforts to detect and delete it. This raises the question if content moderation is simply not effective enough, or if the definitions applied to detect hate speech do not align with users' perceptions of hate speech. We approached this

question combining a representative survey and a mobile longitudinal linkage analysis. This approach enabled us to compare 1) self-reported exposure to hate speech in the survey with observed exposure in the experience sampling study, and 2) subjective exposure from media users' perspective (uploaded screenshots) with objective exposure (content analysis). Further, we tested if social identity biases perceptions of hate speech using an experimental variation of incivil statements in the survey.

According to our survey, about two-thirds of the Swiss population encounter hate speech in their media use, aligning with a cross-national survey (Reichelmann et al., 2021) where 71% of young adults reported exposure to hate speech in the past three months. About one-third of the Swiss population sees hate speech at least once a week. Compared to Swiss adolescents (Külling et al., 2021), adults have a lower exposure frequency. However, study 2 suggests viewing these numbers cautiously. First, less than half of respondents in our mobile longitudinal linkage study uploaded enough content to support their self-reported exposure frequency, indicating potential inaccuracy in self-reports. Although, this finding might be accentuated by the measurement error associated with this type of data collection (Otto et al., 2022). Second, about half of the uploaded screenshots contained impoliteness rather than hate speech, suggesting that it is hard for media users to distinguish between more and less severe forms of uncivil messages. Overall, our studies demonstrate the importance of using multiple approaches to measure exposure to hate speech when evaluating the success of content moderation or other interventions targeting hate speech. While self-reports may overestimate exposure, mobile longitudinal linkage studies may underestimate it. Future research could employ more passive methods like screen recording, website tracking, or data donations to obtain additional exposure indicators – though these methods cannot capture participants subjective evaluations of content as hate speech or not.

Our studies further offer insights into why media users may perceive impolite statements as hate speech. First, statements attacking minorities were more likely evaluated as hate speech, demonstrating the important role of group status for perceptions of hate speech. Moreover, threats seem to be unequivocally considered hate speech, but evaluations vary for insults, defamation, and impoliteness. For these – more contested forms of hate speech from a user perspective – social identity affects the evaluation; users are more likely to consider such statements as hate speech when they identify with the attacked group. This demonstrates that hostile media bias (Vallone et al., 1985) distorts perceptions of uncivil messages, making them seem more severe when one's social identity is attacked. However, the observed effects were very small, and despite many statements being perceived as hate speech, participants do not necessarily believe they should be deleted or the authors reported – except for threats.

Conclusion

In summary, our studies suggest that incivility is prevalent in online media use, but self-reports likely overestimate the exposure to actual hate speech. Media users have a nuanced view and do not consider all forms of hate speech severe enough to necessitate consequences. Identification with the target group increases perceptions of hate speech and the wish for punishment. This leads to two key conclusions: First, impolite

statements contribute strongly to the perceived proliferation of online hate speech but are often seen as legitimate. Therefore, the perceived prevalence of hate speech does not fully reflect its societal impact and is not a reliable indicator for the performance of interventions like content moderation. Second, there seems to be a lack of awareness among the Swiss population about potential legal consequences of offensive and defamatory statements, leading to a higher tolerance of hate speech.

This study contributes to the field of political communication by shedding light on the complex dynamics of online hate speech perception. Our findings emphasize the role of social identity in shaping these perceptions and reveal a discrepancy between academic and public understandings of hate speech. By providing a nuanced understanding of public perceptions of online hate speech, this study offers valuable insights for policy makers, educators, and platforms seeking to combat online hate speech and foster a more respectful and inclusive digital discourse.

Open Scholarship

This article has earned the Center for Open Science badges for Open Data, Open Materials and Preregistered. The data and materials are openly accessible at https://osf.io/qn94h, and https://doi.org/10.17605/OSF.IO/THK6F.

Notes

1. The study was financed by the Swiss Federal Office of Communication (OFCOM). We do not assume that (biased) perceptions of hate speech are country-specific, but the specific content and frequency of exposure may differ across contexts. Thus, some findings may be generalizable while others may not.
2. Link to the preregistration: https://osf.io/thk6f/?view_only=80d131276ac74c0bbbaeaf6586e1b4da
3. See section 'Content analysis: operationalization and reliability' in study 2 for more details about these characteristics and their distinction.
4. Intervista's panel is a probability panel (no self-selection) with the limitation of online access as a necessary precondition to participate in online surveys.
5. This deviates from the preregistration, where we stated that participants scoring 1-5 would be considered as left- and 7-11 as right-wing. In retrospect, we found this not very intuitive and have adapted the categorization. This has some impact on the results. With the more liberal (larger) groups for political orientation, there is no significant difference for the perception of insults (but still for impoliteness) and the desired consequences.
6. Additionally, we estimated Wilcoxon rank-sum tests with weighted data using the R-package survey. The results are the same as for the ANOVAs with unweighted data with significant differences for insults and impoliteness, but not for threats and defamation.
7. Additionally, we estimated Wilcoxon rank-sum tests with weighted data using the R-package survey. The results are the same as for the ANOVAs with unweighted data; there are significant differences for insults and impoliteness, but not for threats and defamation.
8. Study 1 shows that the German and French speaking regions do not differ in terms of exposure vs. non-exposure to hate speech ($F(1,15) = 0.145$, $p = 0.709$) or in terms of frequency of exposure ($F(7,9) = 1.399$, $p = 0.312$).
9. See Swiss Criminal Code (SCC 311.0): https://www.rhf.admin.ch/rhf/de/home/strafrecht/rechtsgrundlagen/national/sr-311-0.html

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Swiss Federal Office of Communications (OFCOM).

Data availability statement

The data that support the findings are available on OSF. The screenshots may contain identifying information, therefore only the coded data but not the screenshots themselves can be shared. The codebook and all questionnaires can also be found on OSF: https://osf.io/qn94h/?view_only = 512c4628fb054c5695beb921244a5624

Notes on contributors

Dominique S. Wirz, PhD University of Zurich, is an Assistant Professor at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam. Her research focuses on media effects in entertainment and political communication.

Sina Blassnig, PhD University of Zurich, is Full Professor at the Department of Communication and Media Research (DCM) at the University of Fribourg and Director of the Institute for Digital Communication and Media Innovation (IDCMI) in cooperation with the University of Applied Sciences of the Grisons (FHGR). Her research interests lie in the areas of digital communication, digital journalism and media innovation, political communication, media use, and media systems in a comparative perspective.

ORCID

Dominique S. Wirz http://orcid.org/0000-0003-2688-8506

References

Ariyanto, A., Hornsey, M. J., & Gallois, C. (2007). Group allegiances and perceptions of media bias: Taking into account both the perceiver and the source. *Group Processes & Intergroup Relations*, *10*(2), 266–279. https://doi.org/10.1177/1368430207074733

Arpan, L. M., & Nabi, R. L. (2011). Exploring anger in the hostile media process. *Effects on News Preferences and Source Evaluation. Journalism & Mass Communication Quarterly*, *88*(1), 5–22. https://doi.org/10.1177/107769901108800101

Bornschier, S., Häusermann, S., Zollinger, D., & Colombo, C. (2021). How "Us" and "them" relates to voting behavior—social structure, social identities, and electoral choice. *Comparative Political Studies*, *0010414021997504*, https://doi.org/10.1177/0010414021997504

Calvert, C. (1997). Hate speech and Its harms: A communication theory perspective. *Journal of Communication*, *47*(1), 4–19. https://doi.org/10.1111/j.1460-2466.1997.tb02690.x

Cohen, E. L., Atwell Seate, A., Kromka, S. M., Sutherland, A., Thomas, M., Skerda, K., & Nicholson, A. (2020). To correct or not to correct? Social identity threats increase willingness to denounce fake news through presumed media influence and hostile media perceptions. *Communication Research Reports*, *37*(5), 263–275. https://doi.org/10.1080/08824096.2020.1841622

Council of Europe. (2022). *Recommendation CM/Rec (2022)16 of the Committee of Ministers to member States on combating hate speech.* https://search.coe.int/cm/Pages/result_details.aspx?ObjectId = 0900001680a67955#globalcontainer

Fortuna, P., & Nunes, S. (2019). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication*, *38*(5), 624–646. https://doi.org/10.1080/10584609.2020.1830322

Hartmann, T., & Tanis, M. (2013). Examining the hostile media effect as an intergroup phenomenon: The role of ingroup identification and status: Hostile media effect. *Journal of Communication*, *63*(3), 535–555. https://doi.org/10.1111/jcom.12031

HateAid (2021). *Grenzenloser Hass im Internet – Dramatische Lage in ganz Europa* [Borderless Hate on the Internet – Dramatic situation in whole Europe]. https://hateaid.org/wp-content/uploads/2022/04/HateAid-Report-2021-DE.pdf

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

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*, *47*(6), 795–814. https://doi.org/10.1177/0093650217699933

Kim, Y., & Hwang, H. (2019). When partisans see media coverage as hostile: The effect of uncivil online comments on hostile media effect. *Media Psychology*, *22*(6), 845–866. https://doi.org/10.1080/15213269.2018.1554492

Kühne, R., & Schemer, C. (2015). The emotional effects of news frames on information processing and opinion formation. *Communication Research*, *42*(3), 387–407. https://doi.org/10.1177/0093650213514599

Külling, C., Waller, G., Suter, L., Bernath, J., Willemse, I., & Süss, D. (2021). *JAMESfocus: Hassrede im Internet*.

Matthes, J. (2013). The affective underpinnings of hostile media perceptions: Exploring the distinct effects of affective and cognitive involvement. *Communication Research*, *40*(3), 360–387. https://doi.org/10.1177/0093650211420255

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication*, *11*(0), Article 0.

Nabi, R. L. (2003). Exploring the framing effects of emotion: Do discrete emotions differentially influence information accessibility, information seeking, and policy preference? *Communication Research*, *30*(2), 224–247. https://doi.org/10.1177/0093650202250881

Otto, L. P., Thomas, F., Glogger, I., & De Vreese, C. H. (2022). Linking media content and survey data in a dynamic and digital media environment – mobile longitudinal linkage analysis. *Digital Journalism*, *10*(1), 200–215. https://doi.org/10.1080/21670811.2021.1890169

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

Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., & Adamkovič, M. (2023). Perception of hate speech by the public and experts: Insights into predictors of the perceived hate speech towards migrants. *Cyberpsychology, Behavior, and Social Networking*, *26*(7), 489–498. https://doi.org/10.1089/cyber.2022.0191

Pedersen, R. T., Petersen, N. B. G., & Thau, M. (2023). *Online abuse of politicians: Experimental evidence on Politicians' own perceptions*. https://doi.org/10.21203/rs.3.rs-3376832/v1

Reichelmann, A., Hawdon, J., Costello, M., Ryan, J., Blaya, C., Llorent, V., Oksanen, A., Räsänen, P., & Zych, I. (2021). Hate knows No boundaries: Online hate in Six nations. *Deviant Behavior*, *42*(9), 1100–1111. https://doi.org/10.1080/01639625.2020.1722337

Reid, S. A. (2012). A self-categorization explanation for the hostile media effect. *Journal of Communication*, *62*(3), 381–399. https://doi.org/10.1111/j.1460-2466.2012.01647.x

Reuters (Hrsg.). (2023). *Reuters institute digital news report 2023*. Reuters Institute.

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, *0093650220921314*, https://doi.org/10.1177/0093650220921314

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, *146144482210911*, https://doi.org/10.1177/14614448221091185

Sellars, A. (2016). *Defining hate speech* (SSRN Scholarly Paper 2882244). https://doi.org/10.2139/ssrn.2882244

Siegel, A. A. (2020). Online hate speech. In N. Persily, & J. A. E. Tucker (Hrsg.), *Social media and democracy: The state of the field, prospects for reform* (pp. 56–88). Cambridge University Press.

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

Stahel, L. (2020). *Status quo und massnahmen zu rassistischer hassrede im internet: Übersicht und empfehlungen*. Eidgenössisches Departement des Innern.

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting incivility and impoliteness in online discussions. *Computational Communication Research*, *2*(1), Article 1.

Vallone, R. P., Ross, L., & Lepper, M. R. (1985). The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of Personality and Social Psychology*, *49*(3), 577–585. https://doi.org/10.1037/0022-3514.49.3.577

Wirth, W., Wettstein, M., Wirz, D. S., Ernst, N., Büchel, F., Schulz, A., Esser, F., Weber, E., Dalmus, C., & Engesser, S. (2019). *Codebook. NCCR democracy module II: The appeal of populist ideas and messages*. https://doi.org/10.17605/OSF.IO/RYX42

Wojatzki, M., Horsmann, T., Gold, D., & Zesch, T. (2018, Oktober 19). *Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments*. Proceedings of the 14th conference on natural language processing. KONVENS 2018, Vienna.

Ziegele, M. (2016). *Nutzerkommentare als anschlusskommunikation*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-12822-7