International Journal of Communication 19(2025), 2084–2103 1932–8036/20250005

From Bystanders to Perpetrators: The Influence of Normative Perceptions and Cognitive Empathy on Online Hate in Korea

MINWOONG CHUNG[1]
Central Michigan University, USA

SEYOUNG LEE
HEEJO KEUM
Sungkyunkwan University, Republic of Korea

Online hate is an increasingly concerning phenomenon in the digital media landscape. This research investigates the process by which observers of online hate transition into active hate perpetrators, focusing on the roles of social norms and cognitive empathy. Longitudinal data collected from a two-wave panel survey in Korea show that frequent observations of online hate heighten the perceived prevalence (descriptive norms) and acceptance (injunctive norms) of such behavior. While cognitive empathy diminished the mediating role of injunctive norms in the relationship between observation and perpetration, it unexpectedly amplified the mediating role of descriptive norms. These findings highlight how the effect of normative perceptions on behavior varies based on individuals' empathy levels. This research provides a foundation for a more systematic understanding of the mechanisms through which observers of online hate become perpetrators and for the development of strategies to combat online hate, fostering a more respectful digital environment.

Keywords: online hate, social norms, cognitive empathy, two-wave data, panel study

Online hate is a growing concern in today's digital media. It typically involves the use of digital platforms to spread spoken or written comments that belittle or stigmatize individuals based on their real or presumed membership in distinct social groups, such as those belonging to specific gender identities or those with different political perspectives (Frischlich, Schatto-Eckrodt, Boberg, & Wintterlin, 2021; Obermaier & Schmuck, 2022). In the landscape of online hate, the majority of participants are observers, with only a small proportion actively engaging in hate speech (Wachs et al., 2019). An intriguing aspect of

Minwoong Chung: chung5m@cmich.edu
Seyoung Lee: gethemane@skku.edu
Heejo Keum (corresponding author): hkeum@skku.edu
Date submitted: 2024-09-06

[1] This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02088387).

this dynamic is the potential for observers to transition into the role of perpetrators (e.g., Wachs et al., 2022). Observers can become active contributors to the cycle of online hate, transforming from passive witnesses to active creators/disseminators of hateful content. However, the existing literature offers limited insights into the mechanisms underlying this transition. The question of why and how observers of online hate become active participants remains largely unanswered, creating a significant gap in our understanding of this phenomenon. This study seeks to address this gap by investigating the factors that motivate observers to engage in online hate and the processes by which they transition into active contributors.

This study examined how observing online hate led to its perpetration, focusing on the mediating roles of social norm perceptions. This study also tested the moderating role of cognitive empathy as an individual trait in the relationship between normative perceptions and behavior. Cognitive empathy—the ability to deliberately comprehend others' emotions and feelings by adopting their perspective—is a key factor in prosocial behaviors. However, surprisingly little evidence is available on how it interacts with social norms. Longitudinal data from a two-wave panel survey showed that frequent observation of online hate predicted a higher perceived prevalence and acceptance of the behavior, which in turn predicted its perpetration. A stronger effect of prevalence perceptions on behavior was found among participants who reported higher levels of cognitive empathy, while a stronger effect of perceived social approval on behavior was observed among participants who exhibited lower levels of it. The findings are expected to offer valuable insights into the mechanism by which observers of online hate can become perpetrators, potentially guiding the development of more effective interventions to curb the impact of online hate.

Online Hate

While hateful communication occurs in both offline and online environments, certain characteristics of online spaces contribute to its greater prevalence. The online disinhibition effect (Suler, 2004) suggests that factors such as anonymity, the absence of immediate physical presence, and a lack of real-time interactions can fuel toxic online behavior. These conditions create a sense of detachment from one's real-world identity and reduce accountability, making individuals more likely to engage in aggressive communication behaviors they might avoid in offline settings. In addition, specific affordances of digital platforms can further facilitate the spread and social endorsement of hateful content. For example, features such as "Like" buttons or upvote mechanisms allow users to express approval with minimal effort (Haim, Kümpel, & Brosius, 2018), sending clear social signals about a post's popularity and acceptance to both content creators and observers (Walther, 2022). These mechanisms can reinforce and amplify hate speech within online communities, further normalizing and encouraging such behavior.

Online hate is a prevalent communication phenomenon to which people are frequently exposed. However, most available data focus on children and young adults, despite its impact across all age groups. For instance, Wachs et al. (2019) conducted a large-scale survey of adolescents (12–18 years) across eight countries (Cyprus, Germany, Greece, India, South Korea, Spain, Thailand, and the United States) and found that nearly half of the participants had witnessed online hate targeting sex, religion, race, or sexual orientation at least once in the past 12 months. Beyond adolescents, a study of individuals aged 18 to 25 across six countries (Finland, France, Poland, Spain, the United Kingdom, and the United States) found that 70.7% of respondents had encountered online hate in the past three months (Reichelmann et al., 2021).

The most frequently targeted groups, listed in order from most to least common, included race/ethnicity, sexual orientation, nationality, gender, religion, political orientation, appearance, and disability.

Online hate is also a prevalent and significant issue in Korea. According to the National Human Rights Commission of Korea (2021), 62% of respondents (1,200 men and women aged 15 and older residing in South Korea) reported experiencing online hate over the past year. The targeted groups of online hate in Korea also exhibit patterns similar to those identified in previous cross-cultural studies, with hate speech often directed toward categories such as gender, region, nationality/race, sexual orientation, religion, political orientation, disability, and age (Lim, Lee, & Keum, 2022). Similarly, Koo, Suh, Chung, Sohn, and Han (2024) analyzed online hate patterns by crawling text data from 11 of the most frequently visited online community platforms in South Korea between 2015 and 2022. Their findings closely aligned with Lim et al.'s (2022) framework and the broader literature on online hate, identifying seven prominent categories: age, politics, race/ethnicity, gender, religion, disability, and sexuality. These categories, ranked from most to least frequent, largely mirrored Lim et al.'s classification, except for the omission of region (Koo et al., 2024).

While much of the literature on online hate perpetration has focused on individual psychological traits, recent research has increasingly highlighted the societal influences that contribute to hate expression (Bernatzky, Costello, & Hawdon, 2022; Bührer, Koban, & Matthes, 2024; Walther, 2022). For instance, Walther (2024) proposes a social approval theory of online hate, arguing that individuals may engage in hate speech not only to express hostility but also to seek validation from like-minded communities. The study suggests that indicators of social approval, such as likes or supportive comments, reinforce and amplify hateful messages. This reinforcement effect suggests that individuals may be more likely to continue or escalate hateful behavior when they perceive social approval from their peers, creating a cycle in which online hate is normalized and perpetuated. In response to the broader call to examine societal influences on online hate, this study explores how social norms shape online hate perpetration, focusing on their interplay with empathy, an individual factor widely recognized for its potential to curb hateful behavior.

The Mediating Role of Social Norms

Mounting empirical evidence indicates that observing online hate can lead to its perpetration. For instance, cross-sectional studies of teenagers and adolescents found a significant, direct positive association between observing online hate and participating in such behavior (Wachs et al., 2019, 2022). Further evidence comes from a study on individuals in their late teens and early 20s in Korea (Chung, Lee, & Keum, 2023), which also pointed to a significant positive relationship between exposure to and perpetration of online hate. This pattern extends beyond young people, as a survey conducted on the general population in Germany (Frischlich et al., 2021) revealed that seeing online hate was significantly and positively related to engaging in similar uncivil behaviors, such as liking, commenting on, and sharing such content.

The theoretical model of norm-building processes (Geber & Hefner, 2019) sheds light on the effect of observing a behavior on conducting that same behavior. The model suggests that witnessing people engage in a particular behavior drives the formation of normative perceptions regarding the behavior, which in turn guides subsequent actions (i.e., normative conformity). The impact of observation in forming

normative perceptions is rooted in social learning theory (Bandura, 1977), which proposes that individuals learn social norms or behavioral guidelines by observing the behavior of others, particularly noting the rewards/punishments resulting from those actions.

Following Cialdini, Reno, and Kallgren's (1990) distinction, Geber and Hefner's (2019) model specifically posits that two types of norms can mediate the relationship between observing a behavior and performing that behavior. First, descriptive norms refer to people's perceptions of what other people are doing. People tend to align with descriptive norms primarily because they offer a heuristic guide to what constitutes suitable and effective behavior, especially in uncertain situations. Second, injunctive norms refer to individuals' perceptions of what behavior is socially approved. Individuals comply with this type of norm to secure social approval or evade social penalties (Cialdini et al., 1990). Meta-analyses of normative influences show that perceived social norms have a small to moderate effect on behavior, with injunctive norms having a stronger influence on behavior than descriptive norms (e.g., McEachan et al., 2016; Rhodes, Shulman, & McClaran, 2020).

Altogether, it is expected that the more often people observe online hate, the more frequently they engage in such behavior. Specifically, frequent observation of online hate can drive people's perceptions that the behavior is widely prevalent (i.e., descriptive norms) and socially approved (i.e., injunctive norms). These normative perceptions can lead to normative conformity, especially in situations of uncertainty or when individuals desire social approval by performing the behavior (Walther, 2022). To test these mediating roles of social norm perceptions in the relationship between witnessing and perpetrating online hate, the following hypotheses were proposed:

H1: Online hate observation will positively predict the perpetration of online hate.

H2a: Perceived descriptive norms will mediate the impact of observing online hate on its perpetration; the observation of online hate will lead to a greater perceived prevalence of this behavior, which will in turn predict more frequent perpetration.

H2b: Perceived injunctive norms will mediate the impact of observing online hate on its perpetration; the observation of online hate will lead to a greater perceived approval of this behavior, which will in turn predict more frequent perpetration.

The Moderating Role of Cognitive Empathy

Theories of social norms suggest that normative influence on behavior is often situational (Chung & Rimal, 2016). Geber and Hefner's (2019) model also supports this idea. Specifically, the model proposes that the relationship between norms and behavior is reinforced when people perceive their actions as observable by important others, particularly when these referent individuals bear similarities with themselves. The expanded theory of normative social behavior (Rimal & Yilma, 2021) offers a more comprehensive perspective on the moderating effects, suggesting three broad categories of moderators: behavioral, individual, and contextual factors. First, behavioral variables correspond to the characteristics of the actions under consideration, such as their addictiveness—that is, the degree to which a certain

behavior leads to compulsive and often harmful attachment. Second, individual factors encompass individual traits like self-efficacy (i.e., an individual's belief in their ability to perform a specific behavior; Bandura, 1977). Third, contextual attributes pertain to social and environmental factors, such as situational ambiguity (i.e., the degree of unfamiliarity an individual experiences in a particular situation).

This study investigates the moderating role of cognitive empathy on the norm-behavior relationship as an individual trait. Broadly, empathy refers to how a person responds to "the observed experiences of another" (Davis, 1983, p. 113), and it encompasses two main dimensions: Cognitive empathy refers to the ability to understand others' emotions and feelings through perspective-taking, while affective empathy involves vicariously experiencing others' emotions (Cuff, Brown, Taylor, & Howat, 2016; Reniers, Corcoran, Drake, Shryane, & Völlm, 2011). There is considerable literature suggesting that empathy, regardless of its dimension, is positively related to various prosocial behaviors (see Yin & Wang, 2023 for a meta-analysis). Notably, studies have shown that empathy is significantly and negatively associated with both the acceptance (Celuch et al., 2022) and perpetration of online hate (Wachs et al., 2022). Furthermore, a longitudinal survey study showed that empathy weakened the effect of observing homophobic insults online on the perpetration of such behavior, as reported in a follow-up survey one year later (Wright & Wachs, 2021).

This study focuses on the role of empathy because of its potential influence on how individuals respond to social norms. That is, the direction and strength of normative influence on behavior can depend on an individual's dispositional empathy, particularly in how it interplays with the cognitive processing of normative cues. When a behavior is perceived as prevalent or socially accepted, some individuals may adopt it at face value, simply conforming to the majority. However, others may first seek to understand the motivations or feelings behind the behavior before deciding whether to follow it. This is especially likely to be relevant when social norms are dysfunctional—that is, when the majority engages in socially undesirable behavior. In such cases, individuals with a stronger tendency to comprehend others' perspectives may be less likely to mindlessly follow negative normative influences and more likely to critically evaluate relevant information before deciding whether to conform.

Although direct evidence is scarce, some indirect evidence suggests that individuals who are motivated to process information thoroughly before making decisions are less susceptible to social influence (Schär, 2021). Additionally, several studies have shown that perspective-taking reduces automatic conformity to negative group behaviors, such as the endorsement of stereotypes and preference for one's own social group (Galinsky & Moskowitz, 2000; Todd, Galinsky, & Bodenhausen, 2012). These findings suggest that individuals with a strong ability to take others' perspectives may be more resistant to harmful social influences and more deliberate in their behavioral choices. Nevertheless, evidence remains scarce on how cognitive empathy influences behavior when people perceive norms as favoring uncivil online communication because of frequent exposure.

To bridge this gap, this study examines how cognitive empathy, which involves perspective-taking skills, moderates the relationship between perceived norms and behavior in the context of online hate. Because this study focuses on how individuals cognitively engage with normative information, cognitive empathy, rather than affective empathy, is the primary focus. Cognitive empathy has been shown to reduce

various forms of online incivility (Barlińska, Szuster, & Winiewski, 2015, 2018) and shape how individuals interpret and respond to perceived norms (Lay, Zagefka, González, Álvarez, & Valdenegro, 2020). Thus, individuals with stronger perspective-taking skills are expected to critically evaluate prevailing norms about online hate, rather than conforming uncritically. Therefore, the following moderated mediation hypotheses were generated:

H3a: *Cognitive empathy will moderate the indirect positive effect of observing online hate on the perpetration of such behavior, through the mediation of perceived descriptive norms. Specifically, the influence of perceived descriptive norms on the perpetration of online hate will weaken as cognitive empathy increases.*

H3b: *Cognitive empathy will moderate the indirect positive effect of observing online hate on the perpetration of such behavior through the mediation of perceived injunctive norms. Specifically, the influence of perceived injunctive norms on the perpetration of online hate will weaken as cognitive empathy increases.*

Method

Participants and Procedure

This data stemmed from a large-scale, multi-wave panel study examining online incivility among late teens and early 20s in Korea. Given the unavailability of a national-level sampling frame, the study adopted a non-probability, quota-sampling approach to ensure that the sample reflected the general Korean population in terms of age, sex, and residential area. A private survey company in Korea distributed e-mail invitations to their research panel for the online surveys.

The initial survey (T1) was conducted in late June 2022. Out of 3,050 eligible participants who accessed the survey, 2,252 respondents completed it, leading to a completion rate of 73.84% (2,000 usable responses). Approximately three months after the conclusion of the first survey, the second wave (T2) was rolled out (all survey questionnaires for the current study were identical). Invitations were e-mailed to those who had completed the first survey. Of the 1,072 eligible participants who accessed the second-wave survey, 1,020 completed it, resulting in a completion rate of 95.15%.

After eliminating eight unusable responses (e.g., those who provided identical answers for all questionnaires), data from 1,012 respondents with no missing data were used for the main analyses. The average age of the participants was 21.27 years (SD = 2.35, range 16 to 24), with 55.83% of them (n = 565) being female and 44.17% (n = 447) being male.

Measures

All measurement items were derived from existing literature but were modified to fit the context of the present study. Unless stated otherwise, all variables were measured using a 10-point Likert scale ranging from 1 = *Strongly Disagree* to 10 = *Strongly Agree*. Higher scores denoted either an increased level

of the variable or stronger agreement with the statement. Table 1 provides the full correlation matrix of the study variables. The full measurement items for the key study variables are available in the Appendix.

Table 1. Correlation Matrix of the Study Variables.

	Variable	1	2	3	4	5	6
1	Online hate observation	--					
2	Online hate perpetration	.13***	--				
3	Perceived descriptive norms	.32***	.13***	--			
4	Perceived injunctive norms	.27***	.14***	.69***	--		
5	Cognitive empathy	.08*	−.13***	.04	−.04	--	
6	Age	.04	.06	.03	.05	−.07*	--
7	Gender	−.01	.26***	−.08*	−.08**	−.06	.04

Note. Gender: 0 = Female, 1 = Male. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Online hate observation (T1; α = .94, *M* = 5.01, *SD* = 2.79) and *online hate perpetration* (T2; α = .94, *M* = 1.89, *SD* = 1.62) were assessed using eight items each (Frischlich et al., 2021; Lim et al., 2022). Participants were asked how often they witnessed online hate and, in the case of perpetration, how often they had engaged in such behavior over the past six months (from 1 = *Never* to 10 = *Always*) across eight categories: gender, region, nationality/race, sexual orientation, religion, political orientation, disability, and age. For each category, a few example phrases (e.g., utilizing offensive slang like "doenjang-nyeo" to target women with verbal assaults) were supplied to clarify for participants what constituted online hate.

Perceived descriptive norms (T2; α = .97, *M* = 5.27, *SD* = 2.82) were measured using five items (Chung & Lapinski, 2019), with an example item being "The majority of people in Korea engage in online attacks on others." *Perceived injunctive norms* (T2; α = .98, *M* = 4.42, *SD* = 2.89) were similarly measured with five items (Chung & Lapinski, 2019; Lapinski, Anderson, Shugart, & Todd, 2014). An illustrative item is, "Most people in Korea in general consider engaging in online attacks on others appropriate."

Cognitive empathy (T2; α = .90, *M* = 6.35, *SD* = 1.72) was assessed using five items from the Interpersonal Reactivity Index (Davis, 1980, 1983), specifically assessing the participants' perspective-taking skills. Example items include: "Before criticizing somebody, I try to imagine how I would feel if I were in their place," and "I sometimes try to understand others better by imagining how things look from their perspectives."

Analysis

We performed a parallel mediation analysis (for H1 and H2) and a moderated parallel mediation analysis (for H3) using PROCESS Models 4 and 14 (Hayes, 2022) in *R*, respectively. The moderated parallel mediation model is described in Figure 1. In both analyses, online hate observation (T1) was introduced as a predictor variable, perceived descriptive and injunctive norms (T2) were entered as parallel mediators, and online hate perpetration (T2) was included as an outcome variable. Participants' gender and age were also included as covariates. We used 10,000 bootstrapped samples to calculate the 95% confidence intervals for the indirect paths, adjusted for bias. Statistical significance was determined by path coefficients with a

p-value of less than .05 (for direct effects) or with 95% confidence intervals that did not include 0 (for indirect effects).

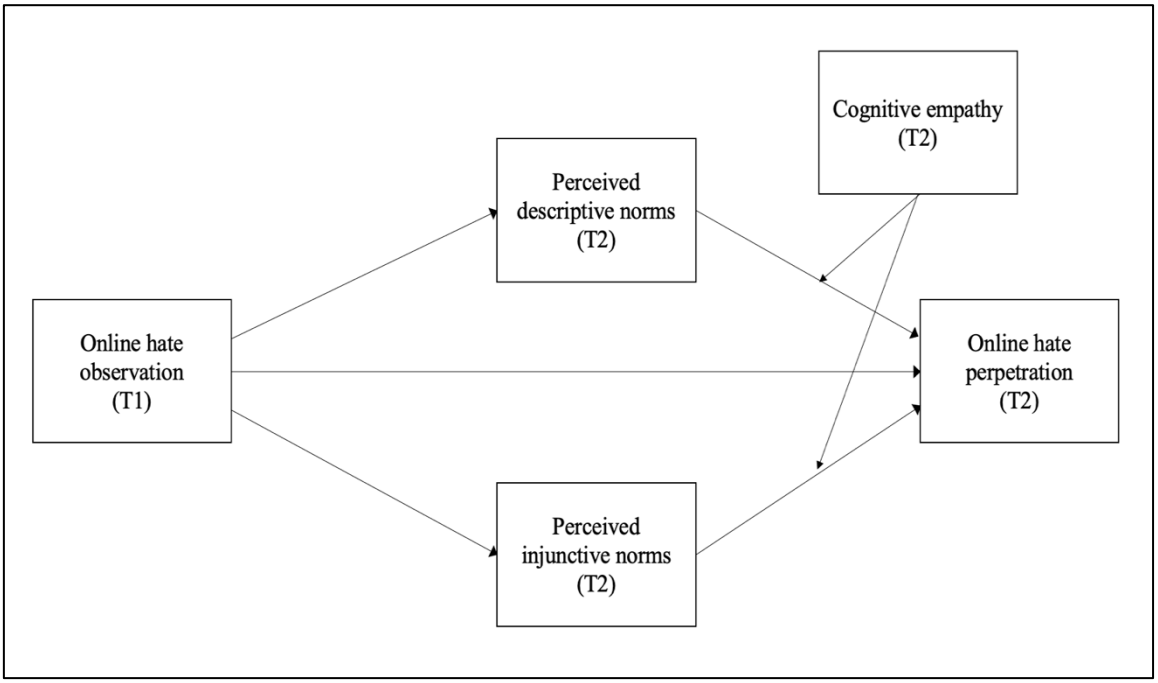

Figure 1. Moderated parallel mediation model of the relationship between online hate observation, social norms, and online hate perpetration, moderated by empathy. T1 indicates Time 1, which refers to the first survey conducted at the initial point of data collection, while T2 represents Time 2, the second survey conducted at a later point in time.

For the moderated mediation test, empathy (T2) was entered as a moderator of the norms-behavior relationship. Continuous variables were mean-centered, and unstandardized beta coefficients were utilized for result interpretation. Where a significant interaction was observed in the data, the conditional indirect effect of mediators on the outcome variable was examined at three levels of each moderator: one standard deviation below the mean (low), the mean (medium), and one standard deviation above the mean (high).

Results

Attrition

Out of the 2,000 individuals who supplied usable data in the initial survey, approximately half (n = 1,012, 50.6%) successfully completed the second-wave survey, notwithstanding repeated attempts to encourage their participation. We conducted a series of t-tests to determine if there was any significant difference in the key study variables between those who dropped out and those who completed both surveys. On average, participants who completed both surveys reported significantly higher perceived descriptive norms at T2, $t(1998) = -3.02$, $p < .01$, Cohen's $d = -.13$, and perceived injunctive norms at T2, t(1998)

= −3.41, $p < .001$, Cohen's $d = -.15$. On the other hand, those participants who dropped out reported a significantly higher frequency of online hate perpetration at T1, $t(1998) = 2.67$, $p < .01$, Cohen's $d = .12$. No significant difference was found in online hate observation and empathy.

Hypotheses Testing

H1 predicted that online hate observation would positively predict the perpetration of online hate. The total effect model showed that the overall model significantly predicted online hate perpetration, $F(3, 1008) = 32.78$, $p < .001$, $R^2 = .09$. The results also revealed that the observation of online hate at T1 significantly and positively predicted the perpetration of the behavior at T2 ($B = 0.08$, $p < .001$). That is, the more people frequently witnessed online hate, the more frequently they engaged in such behavior, as reported three months later. Thus, the data were considered consistent with H1.

H2a anticipated the mediating role of perceived descriptive norms, such that observing online hate would lead to a greater perceived prevalence of the behavior, which in turn would result in more frequent perpetration of online hate. Online hate observation at T1 significantly and positively predicted the perceived prevalence of online hate at T2 ($B = 0.32$, $p < .001$). Nonetheless, the increased prevalence perception at T2 was not significantly associated with online hate perpetration at T2 ($p = .22$). In addition, a bootstrap confidence interval for the indirect effect also included zero ($B = 0.01$, 95% CI = [−0.004, 0.024]), indicating that perceived descriptive norms did not mediate the relationship between the observation and perpetration of online hate. Thus, the data were not consistent with H2a.

H2b predicted the mediating role of perceived injunctive norms: Online hate observation would predict a greater perceived social approval of the behavior, which in turn would lead to more frequent perpetration of online hate. Witnessing online hate at T1 significantly and positively predicted perceived social approval of the behavior at T2 ($B = 0.27$, $p < .001$). In addition, enhanced social approval perception at T2 was significantly and positively linked to the perpetration of online hate at T2 ($B = 0.06$, $p = .02$). The confidence interval for the indirect effect did not include zero ($B = 0.02$, 95% CI = [0.003, 0.029]), suggesting a significant positive mediating role of perceived injunctive norms. Hence, the data supported H2b. Table 2 displays the results of the parallel mediation analysis.

Table 2. Parallel Mediation Model: Total, Direct, and Indirect Effects of Online Hate Observation on Online Hate Perpetration.

	B	*SE*	*t*	95% CI
Total Effect of OBSRV on PERP	0.075	0.018	4.304	[0.041, 0.109]
Direct Effect of OBSRV on PERP	0.050	0.018	2.752	[0.015, 0.086]
	B	*Boot SE*		95% CI
Total Indirect Effect of OBSRV on PERP	0.025	0.006	--	[0.014, 0.037]
Indirect Effect through PDN	0.010	0.007	--	[−0.004, 0.024]
Indirect Effect through PIJN	0.015	0.007	--	[0.003, 0.029]

Note. OBSRV = Online hate observation, PDN = Perceived descriptive norms, PIJN = Perceived injunctive norms, PERP = Online hate perpetration.

H3 posited that cognitive empathy would temper the indirect influence of descriptive (H3a) and injunctive norms (H3b) on behavior. A significant positive interaction effect was observed between perceived descriptive norms at T2 and cognitive empathy at T2 (B = 0.04, p = .01). A moderated mediation analysis revealed that the indirect effect of observing online hate on perpetrating such behavior through perceived descriptive norms was significant only for high levels of cognitive empathy (B = 0.03, 95% CI = [0.010, 0.051]). No significant moderated mediation effect was found for low (B = −0.01, 95% CI = [−0.037, 0.010]) or medium (B = 0.01, 95% CI = [−0.005, 0.023]) cognitive empathy levels. These results suggest that for individuals with high levels of cognitive empathy, perceived descriptive norms positively mediate the relationship between witnessing and perpetrating online hate. In addition, the influence of descriptive norms on online hate perpetration intensified as cognitive empathy increased. Given the unexpected interaction pattern, the data were deemed inconsistent with H3a.

A significant negative interaction effect was found between perceived injunctive norms at T2 and cognitive empathy at T2 (B = −0.03, p = .01). A moderated mediation analysis revealed that the indirect effect of perceived injunctive norms on the outcome variable was significant for low (B = 0.03, 95% CI = [0.011, 0.055]) and medium (B = 0.02, 95% CI = [0.004, 0.030]) levels of cognitive empathy. However, a significant moderated mediation was not found for the high level of cognitive empathy (B = 0.00, 95% CI = [−0.016, 0.017]). These results imply that perceived injunctive norms mediate the influence of observing online hate on perpetrating the behavior, particularly for people with low and medium levels of cognitive empathy. In addition, the influence of injunctive norms on online hate perpetration generally lessens as cognitive empathy increases. Therefore, the data were considered consistent with H3b. Table 3 summarizes the results of the moderated mediation analysis.

Table 3. Moderated Mediation Model: Indirect Effect of Online Hate Observation on Online Hate Perpetration Through Social Norms, Moderated by Empathy.

Mediator	Conditional indirect effects of empathy			
	Condition	Effect/Index	Boot SE	95% CI
Perceived descriptive norms	Low	−0.012	0.012	[−0.037, 0.010]
	Medium	0.008	0.007	[−0.005, 0.023]
	High	0.029	0.011	[0.010, 0.051]
Index of moderated mediation (OBSRV → PDN → PERP)	Cognitive empathy	0.012	0.005	[0.003, 0.023]
Perceived injunctive norms	Low	0.031	0.011	[0.011, 0.055]
	Medium	0.016	0.007	[0.004, 0.030]
	High	0.000	0.008	[−0.016, 0.017]
Index of moderated mediation (OBSRV → PIJN → PERP)	Cognitive empathy	−0.009	0.004	[−0.018, −0.001]

Note. OBSRV = Online hate observation, PDN = Perceived descriptive norms, PIJN = Perceived injunctive norms, PERP = Online hate perpetration. Three levels of each moderator = one standard deviation below the mean ("Low"), the mean ("Medium"), and one standard deviation above the mean ("High").

Discussion

Online hate is a pervasive issue in digital spaces. This study explored the roles of observers, who constitute the majority in online hate contexts, testing how hate observation influences perpetration, with social norms and cognitive empathy playing key roles. Data from a two-wave longitudinal survey demonstrated that observing online hate led to an increased perception of its prevalence and approval. Perceptions of social approval were, in turn, linked to hate perpetration, and the effect of cognitive empathy differed based on the types of social norms. These findings illuminate the mechanisms whereby observing online hate can lead to perpetrating such behavior and identify conditions that can weaken this effect. Longitudinal design offers a dynamic view of these phenomena and strengthens the potential for causal inference, an aspect often missing in previous studies. By establishing the sequence of the variables, this study promotes confidence in the directional influence of observed relationships, paving the way for future research. The implications of these findings, both theoretical and practical, will be discussed in subsequent sections.

More frequent observation of online hate significantly led to more frequent perpetration of such behavior. This is an important finding, as it suggests that the perpetration of online hate is not merely an isolated act but possibly a reactionary behavior triggered by repeated exposure to such behavior. This finding is in line with previous literature (e.g., Wachs et al., 2022). Nonetheless, whereas most existing research has been cross-sectional, the current study provides evidence for the temporal order of the relationship using two-wave longitudinal data. We believe this finding provides an empirical foundation for a more advanced test of causality in the relationship, which could further refine our understanding of how exposure to online hate leads to its perpetration.

Observing online hate influenced individuals' perception of its prevalence (i.e., perceived descriptive norms). However, this perception resulted in behavioral conformity only among those with high cognitive empathy levels. That is, we found no general association between descriptive norms and behavior, and cognitive empathy amplified, rather than weakened, the relationship between descriptive norms and online hate perpetration. Although these findings deviate from our expectations, particularly with the moderating effect being positive rather than negative, they are still partially consistent with existing literature: They underscore that the influence of descriptive norms on behavior is often conditional (Chung & Rimal, 2016) and highlight the significant role of empathy in this relationship (Lay et al., 2020).

One possible explanation for the observed interaction pattern is that perspective-taking toward perpetrators serves as a way to reduce uncertainty in situations where socially undesirable behavior appears to be prevalent. When socially undesirable behavior appears prevalent, stepping into the majority's shoes (the perpetrators, in this case) may be an effective and efficient strategy for understanding ambiguity. In this sense, it is not surprising that the mediating role of descriptive norms was observed only among individuals with a high level of cognitive empathy. Moreover, individuals' perspective-taking toward the majority could lead to normative conformity if they find some similarities with the perpetrators and if the behavior is group-defining, for instance (Tajfel & Turner, 2004). Given that the current study measured participants' dispositional cognitive empathy, we cannot rule out the possibility that the perpetrators, rather than the victims, are the ones being empathized with. Future research involving a specific reference group

for empathy (i.e., whether people empathize with the perpetrators or the victims) is encouraged to better understand the interplay between prevalence perceptions and empathy for socially undesirable behavior.

The observation of online hate shaped the perceptions of social approval of such behavior (i.e., perceived injunctive norms), which in turn predicted the perpetration of online hate. This finding is consistent with our study's predictions and provides empirical evidence supporting the theory that one-way observers of online hate become perpetrators is through perceiving social approval of such behavior (Geber & Hefner, 2019). In a broader context, this aligns with a meta-analysis on social norms, which indicates that injunctive norms exert a greater influence on behavior than descriptive norms (Rhodes et al., 2020).

Additionally, cognitive empathy was found to temper the influence of perceived injunctive norms on behavior. Although it is not entirely clear which group—whether perpetrators or victims—was empathized within the current study, this finding is in line with previous literature suggesting that cognitive empathy can effectively prevent various uncivil online behaviors (e.g., Barlińska et al., 2018). Importantly, the mediating role of perceived injunctive norms was no longer significant among people with strong perspective-taking skills. This implies that cognitive empathy, an individual moderating factor, can not only alter the pattern of the relationship between perceived injunctive norms and behavior, but can also break the link entirely. This is theoretically meaningful, as it demonstrates the substantial role a moderator can play in the norm-behavior relationship.

When considering these findings altogether, the role of cognitive empathy differed across norm types. While cognitive empathy weakened the effect of injunctive norms, it strengthened the effect of descriptive norms in the context of online hate. This may, in part, be due to the different empathy targets associated with each type of norm. When individuals are repeatedly exposed to online hate and come to perceive it as prevalent, they may seek to understand why such negative behavior is widespread, as prevalence offers only a numerical cue rather than an explicit social endorsement. Cognitive empathy, which enables people to understand others' feelings and motivations through perspective-taking, is more likely to be directed toward perpetrators in an attempt to rationalize their actions (e.g., "I want to understand why so many people are doing this.").

In contrast, when individuals perceive online hate as socially accepted, they arguably no longer feel the need to understand the perpetrators because they already serve as a social approval cue. Instead, their cognitive empathy is more likely to shift toward victims (e.g., "I want to understand why they are being targeted"). This aligns with prior research indicating that empathy is not a fixed response but instead varies based on situational influences and personal motivations (Zaki, 2014). Given the limited evidence on the moderating effect of empathy in the relationship between various types of social norms and behavior, further research is strongly encouraged to incorporate dispositional empathy and empathy toward specific groups as individual or contextual moderators when expanding the theories of social norms (e.g., Rimal & Yilma, 2021).

These findings also suggest important guidelines for future research on the social aspects of online hate perpetration, particularly regarding the role of social approval and acceptance in motivating online hate. Individuals may engage in online hate not solely because of personal beliefs or attitudes but also as a way to gain validation from others, as posited by the social approval theory of online hate (Walther, 2022,

2024). When hate is perceived as socially acceptable, who an observer empathizes with (perpetrators or victims) and the extent of that empathy may play a key role in the decision to engage in such behavior. Future research should explore how empathy interacts with social motivations in shaping online hate perpetration, particularly in environments where harmful behaviors are normalized.

From a practical perspective, this research highlights several crucial insights. Recognizing that witnessing online hate can lead to its perpetration, employing various methods to automatically filter and conceal uncivil comments online might be a promising strategy to curb further online incivility (Ayo, Folorunso, Ibharalu, & Osinuga, 2020). Furthermore, considering the evidence that perceptions of social norms can mediate the influence of observation on the perpetration of online hate, it would be beneficial to communicate the fact that such behavior is not the norm, but is rather carried out by a minority and is not socially accepted (e.g., Cialdini et al., 2006). This study also emphasizes the potential value of intervention programs designed to enhance empathy (e.g., Soral, Malinowska, & Bilewicz, 2022), as they could deter observers of online hate from becoming perpetrators themselves. However, it is important to exercise caution: Fostering empathy toward perpetrators might yield undesired consequences, so the focus should be on nurturing empathy toward victims of online incivility.

Limitations

This study has several limitations. First, the use of self-report surveys may have introduced social desirability bias into the participants' responses. Specifically, it is possible that participants' empathy scores were overreported, while online hate perpetration was underreported, resulting in smaller variations in each variable compared to others. To mitigate this potential bias and complement the current study, future research is encouraged to use nonintrusive methods for data collection, such as web mining (e.g., Chau & Xu, 2007), to examine whether online hate observers later become perpetrators. In addition, experimental studies manipulating empathy (e.g., Loon, Bailenson, Zaki, Bostick, & Willer, 2018) and social norms (e.g., Lewis, 2015) are recommended to address the limitations of self-report surveys and to better understand causal effects.

Second, although our study demonstrated the effect of observation on the perpetration of online hate, there might be other unexplored factors influencing this behavior (e.g., online disinhibition; Wright & Wachs, 2021). Adopting an experimental study design could help untangle these causal relationships and provide a more robust understanding of how observing online hate can lead to its perpetration. Alternatively, measuring such potential covariates and statistically controlling for them could serve as another way to address these influences.

Lastly, this study primarily concentrated on the influence of cognitive empathy, as it was expected to relate directly to how individuals cognitively process and interpret normative information. The dual roles of cognitive empathy for descriptive and injunctive norms evidenced in this study highlight the complexity of its influence on online hate. To further unpack these dynamics, we recommend that future research examine the role of affective empathy in the relationship between social norms and behavior in the context of online hate. It is generally anticipated that individuals with moderate to high levels of affective empathy, particularly toward victims, are less likely to engage in uncivil online behaviors (Malecki, Kowal,

Dobrowolska, & Sorokowski, 2021). Investigating these aspects will deepen our understanding of how different types of empathy shape online interactions and advance theoretical perspectives on the societal dimensions of online hate.

Conclusion

The pervasive presence of online hate in today's digital media landscape is a significant concern. This research primarily focused on the role of observers in online hate, particularly their potential transformation into active hate perpetrators. Longitudinal data collected from a two-wave panel survey showed that frequent observation of online hate augmented the perceived prevalence and approval of such behavior. Although perceptions of prevalence did not generally lead to subsequent normative conformity, perceptions of social approval did drive the behavior. Nevertheless, the indirect impact of these normative perceptions on behavior differed based on individuals' levels of cognitive empathy. In particular, while cognitive empathy diminished the mediating role of injunctive norms in the relationship between the observation and perpetration of online hate, it unexpectedly amplified the mediating role of descriptive norms in this relationship. This insight is pivotal for a more systematic understanding of the circumstances and mechanisms through which observers of online hate can become perpetrators. It also aids in the development of strategies to combat online hate, fostering a more respectful digital environment.

References

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of Twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review, 38*, 100311. doi:10.1016/j.cosrev.2020.100311

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

Barlińska, J., Szuster, A., & Winiewski, M. (2015). The role of short- and long-term cognitive empathy activation in preventing cyberbystander reinforcing cyberbullying behavior. *Cyberpsychology, Behavior, and Social Networking, 18*(4), 241–244. doi:10.1089/cyber.2014.0412

Barlińska, J., Szuster, A., & Winiewski, M. (2018). Cyberbullying among adolescent bystanders: Role of affective versus cognitive empathy in increasing prosocial cyberbystander behavior. *Frontiers in Psychology, 9*, 1–13. doi:10.3389/fpsyg.2018.00799

Bernatzky, C., Costello, M., & Hawdon, J. (2022). Who produces online hate?: An examination of the effects of self-control, social structure, & social learning. *American Journal of Criminal Justice, 47*(3), 421–440. doi:10.1007/s12103-020-09597-3

Bührer, S., Koban, K., & Matthes, J. (2024). The WWW of digital hate perpetration: What, who, and why? A scoping review. *Computers in Human Behavior, 159*, 108321. doi:10.1016/j.chb.2024.108321

Celuch, M., Oksanen, A., Räsänen, P., Costello, M., Blaya, C., Zych, I., . . . Hawdon, J. (2022). Factors associated with online hate acceptance: A cross-national six-country study among young adults. *International Journal of Environmental Research and Public Health, 19*(1), 1–13. doi:10.3390/ijerph19010534

Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies, 65*(1), 57–70. doi:10.1016/j.ijhcs.2006.08.009

Chung, A., & Rimal, R. N. (2016). Social norms: A review. *Review of Communication Research, 4*, 1–28. doi:10.12840/issn.2255-4165.2016.04.01.008

Chung, M., & Lapinski, M. K. (2019). Extending the theory of normative social behavior to predict hand-washing among Koreans. *Health Communication, 34*(10), 1120–1129. doi:10.1080/10410236.2018.1461586

Chung, M., Lee, S., & Keum, H. (2023). Seeing is behaving? The role of social norms in the relationship between witnessing and perpetrating online hate. *Communication Reports, 36*(3), 149–161. doi:10.1080/08934215.2023.2230284

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence, 1*(1), 3–15. doi:10.1080/15534510500181459

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015–1026. doi:10.1037/0022-3514.58.6.1015

Cuff, B. M. P., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion Review, 8*(2), 144–153. doi:10.1177/1754073914558466

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10*, 1–19.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113–126. doi:10.1037/0022-3514.44.1.113

Frischlich, L., Schatto-Eckrodt, T., Boberg, S., & Wintterlin, F. (2021). Roots of incivility: How personality, media use, and online experiences shape uncivil participation. *Media and Communication, 9*(1), 195–208. doi:10.17645/mac.v9i1.3360

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology, 78*(4), 708–724. doi:10.1037/0022-3514.78.4.708

Geber, S., & Hefner, D. (2019). Social norms as communicative phenomena: A communication perspective on the theory of normative social behavior. *Studies in Communication, Media, 8*(1), 6–28. doi:10.5771/2192-4007-2019-1-6

Haim, M., Kümpel, A. S., & Brosius, H.-B. (2018). Popularity cues in online media: A review of conceptualizations, operationalizations, and general effects. *SCM Studies in Communication and Media, 7*(2), 186–207. doi:10.5771/2192-4007-2018-2-58

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). New York, NY: Guilford Press.

Koo, J.-W., Suh, C. S., Chung, J. W., Sohn, K.-A., & Han, K. (2024). The hatred of all against all? Evidence from online community platforms in South Korea. *PLoS One, 19*(5), 1–17. doi:10.1371/journal.pone.0300530

Lapinski, M. K., Anderson, J., Shugart, A., & Todd, E. (2014). Social influence in child care centers: A test of the theory of normative social behavior. *Health Communication, 29*(3), 219–232. doi:10.1080/10410236.2012.738322

Lay, S., Zagefka, H., González, R., Álvarez, B., & Valdenegro, D. (2020). Don't forget the group! The importance of social norms and empathy for shaping donation behaviour. *International Journal of Psychology, 55*, 518–531. doi:10.1002/ijop.12626

Lewis, N. (2015). Examining normative influence in persuasive health messages: The moderating role of identification with other parents. *International Journal of Communication, 9*, 3000–3019.

Lim, I., Lee, S., & Keum, H. (2022). A study on the development of scale for online incivility. *Korean Journal of Communication & Information, 116*, 215–249. doi:10.46407/kjci.2022.12.116.215

Loon, A. van, Bailenson, J., Zaki, J., Bostick, J., & Willer, R. (2018). Virtual reality perspective-taking increases cognitive empathy for specific others. *PLoS One, 13*(8), 1–19. doi:10.1371/journal.pone.0202442

Malecki, W. P., Kowal, M., Dobrowolska, M., & Sorokowski, P. (2021). Defining online hating and online haters. *Frontiers in Psychology, 12*, 1–4. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2021.744614

McEachan, R., Taylor, N., Harrison, R., Lawton, R., Gardner, P., & Conner, M. (2016). Meta-analysis of the reasoned action approach (RAA) to understanding health behaviors. *Annals of Behavioral Medicine, 50*(4), 592–612. doi:10.1007/s12160-016-9798-4

National Human Rights Commission of Korea. (2021). *Onlain hyeomopyohyeon insik josa* [*Survey on perceptions of online hate speech*] (Unofficial translation). Retrieved from https://www.humanrights.go.kr/base/board/read?boardManagementNo=17&boardNo=7607189&searchCategory=23&page=8&searchType=total&searchWord=&menuLevel=3&menuNo=115

Obermaier, M., & Schmuck, D. (2022). Youths as targets: Factors of online hate speech victimization among adolescents and young adults. *Journal of Computer-Mediated Communication, 27*(4), 1–11. doi:10.1093/jcmc/zmac012

Reichelmann, A., Hawdon, J., Costello, M., Ryan, J., Blaya, C., Llorent, V., . . . Zych, I. (2021). Hate knows no boundaries: Online hate in six nations. *Deviant Behavior, 42*(9), 1100–1111. doi:10.1080/01639625.2020.1722337

Reniers, R. L. E. P., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment, 93*(1), 84–95. doi:10.1080/00223891.2010.528484

Rhodes, N., Shulman, H. C., & McClaran, N. (2020). Changing norms: A meta-analytic integration of research on social norms appeals. *Human Communication Research, 46*(2–3), 161–191. doi:10.1093/hcr/hqz023

Rimal, R. N., & Yilma, H. (2021). Descriptive, injunctive, and collective norms: An expansion of the theory of normative social behavior (TNSB). *Health Communication, 37*(13), 1573–1580. doi:10.1080/10410236.2021.1902108

Schär, A. (2021). Social pressure on heavy thinkers? The influence of need for cognition on pre-purchase social norm nudges. In *Proceedings of the Twenty-Seventh Americas Conference on Information Systems* (pp. 1–11). Montreal, Canada: Association for Information Systems. Retrieved from https://aisel.aisnet.org/amcis2021/sig_hci/sig_hci/1/?utm_source=aisel.aisnet.org%2Famcis2021%2Fsig_hci%2Fsig_hci%2F1&utm_medium=PDF&utm_campaign=PDFCoverPages

Soral, W., Malinowska, K., & Bilewicz, M. (2022). The role of empathy in reducing hate speech proliferation. Two contact-based interventions in online and off-line settings. *Peace and Conflict: Journal of Peace Psychology, 28*(3), 361–371. doi:10.1037/pac0000602

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*(3), 321–326. doi:10.1089/1094931041291295

Tajfel, H., & Turner, J. C. (2004). The social identity theory of intergroup behavior. In J. T. Jost & J. Sidanius (Eds.), *Key readings in social psychology. Political psychology: Key readings* (pp. 276–293). New York, NY: Psychology Press.

Todd, A. R., Galinsky, A. D., & Bodenhausen, G. V. (2012). Perspective taking undermines stereotype maintenance processes: Evidence from social memory, behavior explanation, and information solicitation. *Social Cognition, 30*(1), 94–108. doi:10.1521/soco.2012.30.1.94

Wachs, S., Bilz, L., Wettstein, A., Wright, M. F., Kansok-Dusche, J., Krause, N., & Ballaschk, C. (2022). Associations between witnessing and perpetrating online hate speech among adolescents: Testing moderation effects of moral disengagement and empathy. *Psychology of Violence, 12*(6), 371–381. doi:10.1037/vio0000422

Wachs, S., Wright, M. F., Sittichai, R., Singh, R., Biswal, R., Kim, E., . . . Maziridou, E. (2019). Associations between witnessing and perpetrating online hate in eight countries: The buffering effects of problem-focused coping. *International Journal of Environmental Research and Public Health, 16*(20), 3992. doi:10.3390/ijerph16203992

Walther, J. B. (2022). Social media and online hate. *Current Opinion in Psychology, 45*, 1–6. doi:10.1016/j.copsyc.2021.12.010

Walther, J. B. (2024). The effects of social approval signals on the production of online hate: A theoretical explication. *Communication Research*, 1–27. Advance online publication. doi:10.1177/00936502241278944

Wright, M. F., & Wachs, S. (2021). Does empathy and toxic online disinhibition moderate the longitudinal association between witnessing and perpetrating homophobic cyberbullying? *International Journal of Bullying Prevention, 3*(1), 66–74. doi:10.1007/s42380-019-00042-6

Yin, Y., & Wang, Y. (2023). Is empathy associated with more prosocial behaviour? A meta-analysis. *Asian Journal of Social Psychology, 26*(1), 3–22. doi:10.1111/ajsp.12537

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin, 140*(6), 1608–1647. doi:10.1037/a0037679

Appendix

Full Measurement Items for Key Study Variables

Online hate observation (perpetration) (Frischlich et al., 2021; Lim et al., 2022)

In the past six months, have you seen (used) derogatory expressions, insults, or slurs toward others online, or engaged in related behaviors, based on the following categories?

1. Gender (e.g., *doenjang-nyeo*, *kimchi-nyeo*, *han-nam*)
2. Regional origin or residence (e.g., *jeolla-dian*, *godam-daegu*)
3. Nationality or race (e.g., *ddong-nama*, *jjanggae*)
4. Gender identity (e.g., *ddongkko-chung*, *byeontae*)
5. Religion (e.g., *gae-dokgyo-in*)
6. Political orientation (e.g., *jwa-bbal*, *su-kkol*)
7. Disability (e.g., *byeongsin*, *aeja*)
8. Age (e.g., *teul-ttak*, *kkondae*, *gubsik-chung*)

Note. The eight categories listed above represent common forms of online hate observed in Korean digital discourse; however, other categories may also exist. Certain derogatory terms outlined in the categories above are commonly used to criticize individuals based on characteristics such as gender, age, and other identity-related factors. For example, *doenjang-nyeo* is a pejorative term for women perceived as prioritizing luxury and superficial lifestyle choices, often associated with frequent visits to expensive cafés or a preference for high-end brands and trendy consumption habits. *Kkondae* is a widely used insult directed at older individuals, particularly those who impose traditional values, act condescendingly, or dismiss younger perspectives. Meanwhile, *gubsik-chung* refers to school-aged individuals in a dismissive or mocking manner, implying immaturity, ignorance, or disruptive online behavior.

Perceived Descriptive Norms (Chung & Lapinski, 2019)

1. Most people in Korea use profanity or harsh language toward others.
2. The majority of people in Korea engage in online attacks on others.
3. Most people in Korea label others' opinions as lies or falsehoods.
4. The majority of people in Korea use derogatory and degrading expressions toward others.
5. Most people in Korea engage in aggressive and threatening nonverbal expressions (e.g., using emoticons, punctuation marks, etc.).

Perceived Injunctive Norms (Chung & Lapinski, 2019; Lapinski et al., 2014)

1. It is generally considered appropriate to use profanity or harsh language toward others.
2. Most people in Korea in general consider engaging in online attacks on others appropriate.
3. The majority of people in Korea in general endorse labeling others' opinions as lies or falsehoods.
4. Most people in Korea in general consider using derogatory and degrading expressions toward others acceptable.

5. The majority of people in Korea in general believe that engaging in aggressive and threatening nonverbal expressions (e.g., using emoticons, punctuation marks, etc.) is acceptable.

Cognitive Empathy (Davis, 1980, 1983)

1. Before criticizing somebody, I try to imagine how I would feel if I were in their place.
2. I sometimes try to understand others better by imagining how things look from their perspective.
3. I believe that there are two sides to every question and try to look at them both.
4. I try to look at everybody's side of a disagreement before I make a decision.
5. When I'm upset at someone, I usually try to "put myself in their shoes" for a while.