Online Hate Speech Among Adolescents: Theory, Research, and Recommendations

Sebastian Wachs, Angela Mazzone,
Anja Schultze-Krumbholz, Michelle F. Wright,
Nicola Döring, Dorothy L. Espelage,
Manuel Gámez-Guadix, and Jun Sung Hong

1 Background

Many societies have become increasingly aware of hate speech in recent years, as it has reached unprecedented levels [1]. Online hate speech (OHS) can cultivate a climate of fear, intolerance, and hatred toward social groups. Moreover, OHS can reinforce discriminatory beliefs and actions, intensifying the oppression and marginalization of the targeted group [2]. Adolescents rely heavily on information and communication technologies (ICT) while dealing with different developmental tasks, including identity exploration, development of autonomy, search for belonging, and formation of romantic relationships [3]. As they seek to establish their sense of identity by affiliating with (online) social groups, adolescents become vulnerable to being targeted by hate groups [4]. Equipping adolescents with the skills they need to deal with this emerging online risk constitutes a significant challenge for researchers, educators, practitioners, and caregivers. This chapter provides an overview of definitional issues, theoretical frameworks, research findings, and empirical research on adolescents' (ages 12–18 years) OHS victimization and perpetration. It concludes with future research directions and recommendations for practitioners.

2 Current State

2.1 A New Definition for an Old Phenomenon

Hatred towards particular groups has existed for a long time in human history. Legal experts, such as Mari Matsuda [5], back in the 1980s, introduced the term *racist speech*, which espouses ethnic inferiority, targets historically oppressed groups, and is hateful and condescending. In subsequent years, the term *racist speech* has been replaced by the more compre-

S. Wachs (✉)
University of Münster,
Nordrhein-Westfalen, Germany
e-mail: swachs@uni-muenster.de

A. Mazzone
University of Surrey, Stag Hill Campus,
Guildford, UK

A. Schultze-Krumbholz
Technische Universität Berlin, Berlin, Germany

M. F. Wright
Indiana State University, Terre Haute, IN, USA

N. Döring
Technische Universität Ilmenau, Thuringia, Germany

D. L. Espelage
The University of North Carolina at Chapel Hill,
Chapel Hill, NC, USA

M. Gámez-Guadix
Autonomous University of Madrid, Madrid, Spain

J. S. Hong
Wayne State University, Detroit, MI, USA

Ewha Womans University, Seoul, South Korea

D. A. Christakis, L. Hale (eds.), *Handbook of Children and Screens*,
https://doi.org/10.1007/978-3-031-69362-5_62

hensive term *hate speech*, referring to all forms of expression that spread, incite, promote, or justify hatred against people based on assigned characteristics, including but not limited to gender identity, sexual orientation, disability status, and religious affiliation. However, the definition of hate speech can vary widely depending on who is historically oppressed, what forms of oppression are deemed unacceptable, and where the lines between free and hate speech are drawn [6]. More recently, the term has been extended to the online context and groups of people who are not traditionally oppressed per se (e.g., politicians, journalists, etc.) but are perceived as allies. In addition, the terms hate speech and cyberhate are often used interchangeably, which adds to the inconsistencies regarding terminology and obscures the fact that hate speech occurs online and offline.

A major challenge in current research is that no generally agreed-upon definition exists [6]. To address this gap, Kansok-Dusche et al. [7] conducted a systematic review of definitions in existing online and offline hate speech research conducted with young people and derived the following definition:

> Hate speech is a derogatory expression (e.g., words, posts, text messages, images, and videos) about people (directly or vicariously) on the basis of assigned group characteristics (e.g., ethnicity, nationality, gender, sexual orientation, disability, and religion). Hate speech is based on an intention to harm and it has the potential to cause harm on multiple different levels (e.g., individual, communal, and societal) [7]. (p. 11)

The proposed definition consists of four key elements. First, it encompasses various human behaviors in online and offline settings. Second, it involves targeting people based on assigned group characteristics; however, it is intentionally broad, as it acknowledges that social categories beyond currently marginalized groups could become victims of hate speech. Third, it recognizes that derogatory expressions can cause harm on various levels. Finally, the proposed definition focuses on the intention to harm rather than being limited to biased attitudes or emotions. Despite the effort to systematize existing knowledge, this definition presents challenges, especially in assessment, as it is difficult to assess individuals' intentions behind the observable aggressive speech and the impact on victims, communities, or societies.

2.2 Frequency Rates and Assessment of Online Hate Speech Involvement

According to a recent systematic review, frequency rates for witnessing OHS vary between 31.4% and 68.5%, for perpetration between 4.2% and 32.2%, and victimization between 9% and 23.4% [7]. The varying estimates of frequency rates across different studies can be attributed to country differences (e.g., how hate speech is defined in each particular country), methodological differences, including the reference period (e.g., lifetime, last 3 months), response options (dichotomous or polytomous), whether a definition of OHS is provided beforehand or not, sample characteristics (e.g., age, gender, ethnicity distribution), and whether OHS is measured in general or targeting a specific group (e.g., racist OHS). From a methodological point of view, instruments to investigate young people's involvement in OHS are often based on single-item measures (e.g., In the past 12 months, how often have you witnessed online hate speech?), sometimes with a definition as an introduction to the single-item measures. Using single-item measures is problematic and can lead to limited reliability and validity, as they may not fully capture complex constructs such as OHS or the variability in respondents' perspectives. For example, OHS can alternate between clearly recognizable calls for violence and denigration and more subtle forms (e.g., disguised as irony, offensive jokes, use of stereotypes, and generalizations). Additionally, single-item measures can be more susceptible to measurement error and bias, potentially compromising the accuracy and interpretiveness of findings. Given this complexity, using multiple-item scales to measure various OHS manifestations is critical.

2.3 Theory and Research on Online Hate Speech Perpetration

Several theories have been tested to understand why adolescents share, publish, or produce hateful online content. For example, using the *Online Disinhibition Effect* [10], empirical evidence revealed that toxic online disinhibition was positively linked to OHS perpetration [11]. Applying the *Social Cognitive Theory of Morality* [12], research revealed that a series of moral disengagement mechanisms (i.e., socio-cognitive processes aimed to justify immoral behaviors through attributing blame to the target, dehumanizing the victim, and minimizing agency) are associated with OHS perpetration [13]. Past research showed that the positive association between witnessing and perpetrating OHS was stronger at higher levels of moral disengagement and weaker when moral disengagement was low [13]. In other research, the *Problem Behavior Theory* [14] has been used to conceptualize adolescents' engagement in OHS as a facet of problematic behavior that is interrelated with other problematic behaviors, which all come from an underlying cause or causes, such as certain personality traits (e.g., impulsivity and sensation-seeking). These underlying causes have consistently been found to increase susceptibility to engaging in risky behaviors (e.g., violence, and delinquency). Hate speech perpetration has also been found to be associated with other risk factors such as contact with strangers online, excessive Internet use, and cyberbullying perpetration [15, 16]. Consistent with the *Social Dominance Theory* [17], the persistence of discrimination and prejudice in societies can be attributed to the intersection of ideologies, institutional practices, social dynamics, and personal attributes, reinforced by ideologies that posit certain groups as superior and others as inferior. OHS likely has its foundation in in-group and out-group identification. Extant research has found that adolescents who perceived their in-group as superior were more likely to perpetrate OHS against out-group members [18].

2.4 Theory and Research on Online Hate Speech Victimization

Studies investigating factors leading to OHS victimization often apply the *routine activity theory* [19] as a theoretical framework. According to this theory, adolescents are more likely to be victims when there is a convergence of three factors: Exposure to a motivated offender, a suitable target, and the absence of a capable guardian. Regarding exposure to a motivated offender, research findings suggest that witnessing hate speech, contact with strangers online, deliberate searches of hate-related materials online, hate speech perpetration, and excessive ICT use are linked to hate speech victimization [16, 20, 21]. Regarding target suitability, the research found that individual characteristics (e.g., being female, being gay, having a migration background, and being a member of a minority religion) increased the risk of OHS victimization [22]. Moreover, expressing online support for the LGBTQIA+ community, high disclosure of private information online, offline OHS victimization, low digital media literacy, and experiences of data misuse online increased OHS victimization risk among adolescents [8, 16, 20–22].

Research on the lack of capable guardianship revealed that parental behavior plays a significant role. For example, parents sharing personal information about their children online could increase their children's risk of OHS victimization [16]. Also, parental mediation of children's ICT use (interactions parents have with their children about media use) is relevant to consider. Instructive parental mediation of children's online activities was found to be associated with less hate speech victimization, while restrictive parental mediation was positively related to greater OHS victimization [21]. Parents who adopt instructive mediation might engage in discussions with their children regarding ICT use and its potential risks. In turn, this may result in their children being better educated about the dangers of online interactions and greater compliance with safety recommendations. Conversely, parents adopting restrictive mediation could potentially harm their children's ability to manage problematic online situations. Additionally, these restrictive strategies could be viewed as a threat to children's independence, leading to increased psychological reactance and children's not disclosing their experiences online.

Another avenue of research has focused on the consequences of OHS victimization. For example, victims of OHS experience adverse mental health outcomes, including lower mental well-being and higher anxiety levels, depressive symptoms, fear and insecurity, and sleeping disorders [23–26]. OHS victimization can also impact adolescents' behavior, such as increasing physical aggression, rule-breaking behaviors, and poor academic outcomes (i.e., academic motivation) [27]. In addition, frequently experiencing racist OHS hindered Black adolescents' development of social skills such as empathy, suggesting that OHS victimization can impede adolescents' ability to demonstrate their full potential [28].

Only a few studies have investigated variables that buffer the adverse effects of OHS on victims. For example, one study found that resilience measured individual factors (e.g., social competence, personal competence, and structured style), familial factors (e.g., family cohesion), and a supportive environment outside the family (e.g., social resources), buffered against the effects of OHS victimization on depressive symptoms [25]. Another study revealed that African American adolescents with higher self-esteem and positive ethnic identity experienced less anxiety resulting from racial OHS victimization [29]. This suggests that having a strong sense of self and ethnic identity can buffer against the adverse effects of racist OHS.

3 Future Research

Below are three key questions that we feel OHS scholars need to address over the coming years.

3.1 What Are the Methodological Challenges in Online Hate Speech Research?

As research on OHS among adolescents is at an early stage, there are many pressing challenges to conducting research in this area, including how OHS is defined. While systematically reviewing existing literature might contribute to elaborating a scientifically sound definition, a bottom-up approach involving key stakeholders, including young people, educators, and school personnel, may assist in testing whether existing definitions reflect their lived experiences. Another Achilles' heel of OHS research related to the definition is how OHS is measured. Accurately assessing OHS through research is essential to advancing the research field, evaluating interventions, and informing policymakers. As mentioned above, most research is currently based on single items. Assessing hate speech is further complicated by deciding whether to measure hate speech in general or measure hate speech experienced by or directed at specific target groups (e.g., Muslims). Further, researchers must decide which derogatory expressions (e.g., words, posts, messages, memes, and videos) and which modes (e.g., offensive jokes, use of stereotypes, and generalizations) are captured in their measures.

Furthermore, most research on hate speech among adolescents is based on cross-sectional study designs, which does not allow for establishing temporal associations between OHS and relevant outcomes. Longitudinal and experimental OHS research is needed to refine our descriptive understanding of OHS and increase our knowledge of risk factors and consequences. Finally, there is a lack of innovative data-collection techniques in OHS research among adolescents. Although using peer nominations poses several ethical issues [30], this method might elucidate the social dynamics of OHS. Another innovative approach might be using experience sampling methods (e.g., daily diary) to study "in real time" the daily life of adolescents involved in hate speech and its impact on outcomes concurrently and temporally. Using this technique would allow researchers to understand the impact of memory biases, enhance real-life relevance, and evaluate hypotheses between- and within-person levels [31].

3.2 How Can We Increase Adolescents' Engagement Against Online Hate Speech?

Despite the increase in research focused on hate speech experiences among adolescents, studies to date have mainly focused on perpetrators and victims and have only recently recognized that hate speech can involve others. Adolescents encountering OHS can show moral courage by countering OHS (counterspeech). Counterspeech is defined as a form of citizen-based response to hateful content to discourage it, stop it, or provide support for the victim by, for example, pointing out logical flaws in the hateful content or using facts to counteract misinformation [32]. Until recently, little is known about the factors that increase adolescents' potential or actual engagement in counterspeech, factors preventing them from doing so, and how we can support adolescents to effectively stand up against OHS without putting themselves in danger. Such research should also investigate factors that moderate and mediate the association between predictors and counterspeech to identify the conditions and mechanisms that increase the likelihood of counterspeech.

3.3 What Are Effective Strategies to Prevent Online Hate Speech Involvement Among Adolescents?

At present, evidence-based prevention programs to prevent OHS among adolescents are scarce. Common methods for preventing biased attitudes and promoting positive intergroup relations often involve one or more of the following components: Interventions that encourage intergroup contact (e.g., youth exchange programs or reading materials about members of marginalized groups), knowledge-based interventions (e.g., providing information about minorities and democratic values), and individual skill acquisition (e.g., empathy training). More research is needed to understand the most effective approach to address OHS and whether varying approaches might be more or less effective for different groups of young people. In fact, a multicomponent approach might be effective in tackling OHS. For example, the "HateLess. Together against Hatred" prevention program combines these elements. An evaluation study found that HateLess effectively increases adolescents' empathy for victims, self-efficacy toward intervening, and engagement in counterspeech [33, 34]. More prevention research needs to be conducted to increase the acceptability, fidelity, and sustainability of the existing programs to improve adolescent hate speech-related outcomes. In addition, more research is needed to understand the cross-cultural validity of existing programs and the most effective ways to prepare adolescents for living harmoniously in diverse societies.

4 Recommendations

Some key recommendations from existing research include:

- Raise awareness around the harmful nature of—online and offline—hate speech for individuals and societies to prevent trivialization and justification of perpetrators' behavior.
- Emphasize morality training that aims to raise awareness of the socio-cognitive processes that adolescents might activate to reduce guilt and remorse when perpetrating OHS.
- Encourage civic engagement by offering human rights education and promoting knowledge (e.g., regarding equality, inclusivity, and diversity), attitudes, opportunities, and social-emotional skills (e.g., expressing opinions appropriately).
- Provide cybersecurity and cyber protection information and combine them with behavioral

components to enable adolescents to protect their data and information online.

- Identify and promote young people's social and personal resources that bolster resilience and mitigate adverse effects of OHS victimization.
- Inform parents of effective parental mediation strategies and ensure children's fundamental rights (e.g., informational self-determination and age-appropriate online privacy) without being intrusive.
- Encourage educators and parents to talk regularly and openly with their children about their online experiences.
- Implement digital literacy interventions for young people, teachers, and parents and combine them with ethical and civic courage components to address prejudice, stereotypes, and self-efficacy.
- Consult stakeholders (e.g., adolescents, educators, parents, and social media providers) to design effective policies and provider-based intervention strategies, such as human- and artificial intelligence-based content and comment moderation.

Conflict of Interest and Funding Disclosures None.

References

1. Costello M, Hawdon J. Hate speech in online spaces. In: The Palgrave handbook of international cybercrime and cyberdeviance. Palgrave Macmillan; 2020. p. 1397–416.
2. Soral W, Bilewicz M, Winiewski M. Exposure to hate speech increases prejudice through desensitization. Aggress Behav. 2018;44(2):136–46.
3. Borca G, Bina M, Keller PS, Gilbert LR, Begotti T. Internet use and developmental tasks: Adolescents' point of view. Comput Hum Behav. 2015;52:49–58.
4. Bauman S, Perry VM, Wachs S. The rising threat of cyberhate for young people around the globe. An ecological perspective. Academic Press; 2021. p. 149–75.
5. Matsuda MJ. Public response to racist speech: considering the victim's story. Mich Rev. 1988;87:2320–81.
6. Brown A. What is hate speech? Part 1: the myth of hate. Law Philos. 2017;36:419–68.
7. Kansok-Dusche J, Ballaschk C, Krause N, et al. A systematic review on hate speech among children and adolescents: definitions, prevalence, and overlap with related phenomena. Trauma Viol Abuse Published online. 2022;24:2598.
8. Castellanos M, Wettstein A, Wachs S, et al. Hate speech in adolescents. A binational study on prevalence and demographic differences. Front Educ. 2023;8:1076249.
9. Döring N, Mohseni MR. Fail videos and related video comments on YouTube: a case of sexualization of women and gendered hate speech? Commun Res Rep. 2019;36(3):254–64.
10. Suler J. The online disinhibition effect. Cyberpsychol Behav. 2004;7(3):321–6.
11. Wachs S, Wright MF. The moderation of online disinhibition and sex on the relationship between online hate victimization and perpetration. Cyberpsychol Behav Soc Netw. 2019;22(5):300–6.
12. Bandura A, Barbaranelli C, Caprara GV, Pastorelli C. Mechanisms of moral disengagement in the exercise of moral agency. J Pers Soc Psychol. 1996;71(2):364.
13. Wachs S, Bilz L, Wettstein A, et al. Associations between witnessing and perpetrating online hate speech among adolescents: testing moderation effects of moral disengagement and empathy. Psychol Violence. 2022;12(6):371–81.
14. Jessor R, Jessor SL. Problem behavior and psychosocial development: a longitudinal study of youth. New York: Academic Press; 1977.
15. Bedrosova M, Machackova H, Šerek J, Smahel D, Blaya C. The relation between the cyberhate and cyberbullying experiences of adolescents in The Czech Republic, Poland, and Slovakia. Comput Hum Behav. 2022;126:107013.
16. Wachs S, Mazzone A, Milosevic T, et al. Online correlates of cyberhate involvement among young people from ten European countries: an application of the Routine Activity and Problem Behaviour Theory. Comput Hum Behav. 2021;123:106872.
17. Sidanius J, Pratto F. Social dominance theory. In: Handbook of theories of social psychology. Sage Publications; 2012. p. 418–39.
18. Ballaschk C, Wachs S, Krause N, et al. Dann machen halt alle mit.Eine qualitative Studie zu Beweggründen und Motiven für Hatespeech unter Schüler* innen. Diskurs Kindh- Jugendforsch J Child Adolesc Res. 2021;16(4):463–80.
19. Cohen LE, Felson M. Social change and crime rate trends: a routine activity approach. Am Sociol Rev. 1979;44(4):588–608.
20. Räsänen P, Hawdon J, Holkeri E, Keipi T, Näsi M, Oksanen A. Targets of online hate: examining determinants of victimization among young Finnish Facebook users. Violence Vict. 2016;31(4):708–25.
21. Wachs S, Costello M, Wright MF, et al. "DNT LET' EM H8 U!": applying the routine activity framework to understand cyberhate victimization among adolescents across eight countries. Comput Educ. 2021;160:104026.
22. Obermaier M, Schmuck D. Youths as targets: factors of online hate speech victimization among

adolescents and young adults. J Comput-Mediat Commun. 2022;27(4):zmac012.

23. Gámez Guadix M, Wachs S, Wright M. " Haters back off!" psychometric properties of the coping with cyberhate questionnaire and relationship with well-being in Spanish adolescents. Psicothema. 2020;32(4):567–74.
24. Krause N, Ballaschk C, Schulze-Reichelt F, et al. "Ich lass mich da nicht klein machen!"Eine qualitative Studie zur Bewältigung von Hatespeech durch Schüler/innen. Z Für Bild Published online. 2021;11:169–85.
25. Wachs S, Gámez-Guadix M, Wright MF. Online hate speech victimization and depressive symptoms among adolescents: the protective role of resilience. Cyberpsychology Behav Soc Netw. 2022;25(7):416–23.
26. Tynes BM, Giang MT, Williams DR, Thompson GN. Online racial discrimination and psychological adjustment among adolescents. J Adolesc Health. 2008;43(6):565–9.
27. Tynes BM, Rose CA, Hiss S, Umaña-Taylor AJ, Mitchell K, Williams D. Virtual environments, online racial discrimination, and adjustment among a diverse, school-based sample of adolescents. Int J Gam Comput-Mediat Simul IJGCMS. 2014;6(3):1–16.
28. Lozada FT, Tynes BM. Longitudinal effects of online experiences on empathy among African American adolescents. J Appl Dev Psychol. 2017;52:181–90.
29. Tynes BM, Umana-Taylor AJ, Rose CA, Lin J, Anderson CJ. Online racial discrimination and the protective function of ethnic identity and self-esteem for African American adolescents. Dev Psychol. 2012;48(2):343–55.
30. Cillessen AH, Marks PE. Methodological choices in peer nomination research. New Dir Child Adolesc Dev. 2017;2017(157):21–44.
31. Mölsä ME, Lax M, Korhonen J, Gumpel TP, Söderberg P. The experience sampling method in monitoring social interactions among children and adolescents in school: a systematic literature review. Front Psychol Published online. 2022;13:1504.
32. Garland J, Ghazi-Zahedi K, Young JG, Hébert-Dufresne L, Galesic M. Impact and dynamics of hate and counter-speech online. EPJ Data Sci. 2022;11(3):3.
33. Wachs S, Krause N, Wright MF, Gámez-Guadix M. Effects of the prevention program "HateLess. Together against Hatred" on Adolescents' empathy, self-efficacy, and countering hate speech. J Youth Adolesc. 2023;52:1115–28.
34. Wachs S, Wright MF & Gámez-Guadix M. From hate speech to HateLess. The effectiveness of a prevention program on adolescents' online hate speech involvement. Computers in Human Behavior, 2024;157:108250. https://doi.org/10.1016/j.chb.2024.108250