



OPEN ACCESS

EDITED BY

Peinado Rodriguez Matilde,
University of Jaén, Spain

REVIEWED BY

Jie Liu,
University of Jinan, China
Vera Titkova,
HSE University, Russia

*CORRESPONDENCE

Sebastian Wachs
✉ swachs@uni-muenster.de

RECEIVED 18 February 2026

REVISED 11 May 2026

ACCEPTED 13 May 2026

PUBLISHED 02 June 2026

CITATION

Wachs S, Kansok-Dusche J, Hong JS, Bilz L, Wettstein A and Gámez-Guadix M (2026) Affective, cognitive, and normative correlates of bystander profiles in racist hate speech: cross-level moderation of empathy by anti-hate speech norms. *Front. Educ.* 11:1813373. doi: 10.3389/feduc.2026.1813373

COPYRIGHT

© 2026 Wachs, Kansok-Dusche, Hong, Bilz, Wettstein and Gámez-Guadix. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Affective, cognitive, and normative correlates of bystander profiles in racist hate speech: cross-level moderation of empathy by anti-hate speech norms

Sebastian Wachs^{1*}, Julia Kansok-Dusche², Jun Sung Hong^{3,4}, Ludwig Bilz², Alexander Wettstein⁵ and Manuel Gámez-Guadix⁶

¹Institute of Education, University of Münster, Münster, Germany, ²Department of Health Sciences, Brandenburg University of Technology Cottbus-Senftenberg, Senftenberg, Germany, ³School of Social Work, Wayne State University, Detroit, MI, United States, ⁴Department of Social Welfare, Ewha Womans University, Seoul, Republic of Korea, ⁵Department of Research and Development, University of Teacher Education Bern, Bern, Switzerland, ⁶Biological and Health Psychology Department, Autonomous University of Madrid, Madrid, Spain

Background: Hate speech in schools threatens inclusive classroom climates, positioning bystanders as central actors in either reinforcing or counteracting discriminatory interactions. While prior research has identified individual predictors of bystander behavior, less is known about how affective, cognitive, and normative processes across ecological levels jointly shape heterogeneous response patterns. Grounded in the Social Cognitive Theory, this study examines affective (empathy), cognitive (moral disengagement), and normative (injunctive anti-hate speech norms) correlates of latent bystander profiles at both the student- and classroom-level, and tests whether shared classroom norms condition the association between individual empathy and distinct bystander responses.

Methods: Using a multilevel, person-centered approach, data from 3,225 seventh- to ninth-grade students were analyzed to examine student- and classroom-level correlates of profile membership. Participants completed self-report questionnaires measuring bystander responses empathy, moral disengagement, and injunctive anti-hate speech norms.

Results: At the student-level, higher empathy was positively associated with defender membership but also with revenger and contributor profiles, whereas moral disengagement was linked to a lower likelihood of defending and a higher likelihood of contributing. Stronger perceived injunctive anti-hate speech norms at the student-level were associated with a higher likelihood of defender membership and a lower likelihood of contributor membership. At the classroom-level, stronger injunctive anti-hate speech norms were related to a reduced likelihood of contributor membership. Moreover, classroom injunctive anti-hate speech norms moderated the association between student empathy and defending, such that empathy was less predictive of defender membership in classrooms with stronger anti-hate speech norms.

Conclusion: The findings suggest that adolescents' bystander responses to hate speech reflect an interplay between individual dispositions and classroom normative climates. In particular, classroom injunctive norms appear to moderate the extent to which individual empathy is associated with defending, although this cross-level effect should be interpreted with caution. The study provides a focused multilevel extension of bystander research by highlighting the context-dependent role of empathy.

KEYWORDS

bystanders, empathy, hate speech, moral disengagement, multilevel

1 Introduction

Hate speech has become a pressing concern in school settings, shaping peer interactions and classroom climates (Lehman and Zhang, 2025). By targeting students based on socially ascribed characteristics, hate speech fosters fear and exclusion, reinforcing discriminatory dynamics that undermine inclusive learning environments and schools' commitments to equity and democratic coexistence (UNESCO, 2023). When hate speech occurs, it rarely unfolds in private; peers are typically present as bystanders whose responses range from active defending to passive disengagement or reinforcement of the aggressor (Wachs et al., 2024b). These dynamics are particularly consequential during adolescence—a developmental period marked by heightened sensitivity to peer relationships, identity development, and social norms (Hay et al., 2004). As peer approval becomes a central reference point, hate speech incidents may serve as cues for interpreting social expectations and positioning oneself within peer hierarchies. Understanding why adolescents differ in their bystander responses is therefore critical not only for prevention efforts but also for explaining how hate speech is reproduced or challenged in classroom contexts.

Previous research suggests that bystander behavior in hate speech situations is shaped by both individual dispositions (e.g., empathy, moral disengagement) and contextual factors (e.g., peer relationships) (Ieracitano et al., 2024; Jiménez-Díaz and Del Rey, 2025; Rudnicki et al., 2023). Yet less is known about how classroom-level normative environments shape bystander response patterns. Addressing this gap is essential for understanding whether bystander responses primarily reflect stable individual differences or are contingent on the normative structure of the classroom context.

Drawing on Bandura's (1986) Social Cognitive Theory, which emphasizes reciprocal interactions among personal characteristics, moral self-processes, and social environments, we conceptualize bystander responses to hate speech as shaped by affective (empathy), cognitive (moral disengagement), and normative (injunctive anti-hate speech norms) processes operating at both the student and classroom levels. Building on previously identified latent bystander response profiles (Wachs et al., 2024b), the present study applies a multilevel framework to examine how these processes are associated with differentiated response patterns. By testing cross-level interactions, it further examines how shared classroom norms shape the role of individual empathy in bystander responses to racist hate speech.

1.1 Bystander responses to hate speech in schools

Hate speech refers to harmful and demeaning expressions directed at individuals or groups based on socially ascribed characteristics (e.g., ethnicity, religion, sexual orientation), with consequences that unfold at the individual, social, and societal levels (Kansok-Dusche et al., 2023). Importantly, hate speech as a form of bias-based aggression differs from general (peer) aggression in that it is inherently group-based and rooted in social identity processes (Bedrosova et al., 2022; Palmer and Abbott, 2018). Rather than targeting individuals for idiosyncratic reasons, racist hate speech reinforces group hierarchies and communicates exclusion and denigration based on group membership (Kansok-Dusche et al., 2023). This intergroup dimension has important implications for bystander behavior. Intervening in such situations may not only reflect prosocial motivation but also involve positioning oneself relative to salient group norms, potentially entailing social risks such as exclusion or status loss. Consistent with this reasoning, prior research suggests that bystander responses in bias-based aggression contexts are more strongly shaped by perceived norms and intergroup dynamics than in non-bias-based aggression contexts (Gönültaş and Mulvey, 2023; Pabian, 2025; Palmer and Abbott, 2018). Consequently, shared classroom norms may play a particularly central role in shaping bystander responses in racist hate speech contexts, as they signal general expectations and group-relevant boundaries of acceptable conduct.

Hate speech in school contexts typically occurs in the presence of peers (Castellanos et al., 2023; Krause et al., 2021). Consequently, many students must decide how to respond to hate speech incidents. These responses are consequential, as they may either reinforce exclusionary dynamics or foster protection and inclusion of targeted peers (Wachs et al., 2024a). Research consistently showed that students' bystander responses to hate speech are heterogeneous. For example, Wachs et al. (2024a) identified multiple responses, including comforting the victim, seeking support, actively countering hate speech, ignoring the incident, experiencing helplessness, retaliating against the person who perpetrates hate speech, or reinforcing it. Although prosocial responses are most frequently reported, disengaged and harmful responses remain common, underscoring the need for approaches that capture the full range of adolescents' responses. Much of the existing research has relied on variable-centered approaches that examine single bystander behaviors in isolation. While these approaches generated important insights, they may overlook adolescents' tendency to endorse multiple,

potentially conflicting reactions within the same situation. For instance, defending behaviors may coexist with retaliatory impulses, reflecting qualitatively distinct response patterns rather than independent behavioral tendencies. Person-centered approaches, such as latent profile analysis, address this limitation by identifying systematic configurations of responses. This perspective is particularly well-suited to morally and emotionally complex situations, such as hate speech, where reactions are unlikely to be uniform or mutually exclusive.

Person-centered research supports this perspective by identifying distinct bystander response profiles in school-based hate speech contexts. Research by Wachs et al. (2024b) identified four latent profiles that differed systematically in their patterns of prosocial, antisocial, and avoidant responses. Passive bystanders (34.2%) were characterized by pronounced tendencies toward ignoring and helplessness, combined with below-average engagement in both prosocial and overtly antisocial behaviors. Defenders, the largest group (47.3%), displayed a clearly prosocial pattern, including elevated comforting, countering, and help-seeking behaviors, alongside low endorsement of antisocial and avoidant responses. Revengers (9.8%) were primarily distinguished by heightened retaliatory responses toward the perpetrator and comparatively low engagement in prosocial strategies. Contributors with the smallest profile (8.6%) exhibited a predominantly antisocial configuration, combining reinforcement of the perpetrator with low levels of prosocial intervention and greater disengagement. These profiles illustrate that adolescents' responses to hate speech reflect structured patterns of engagement rather than isolated behavioral tendencies.

1.2 Individual and classroom correlates of latent bystander profiles

From a social cognitive perspective (Bandura, 1986), adolescents' bystander responses to hate speech emerge from interactions among personal characteristics, moral-cognitive processes, and social environments. Guided by this framework, we conceptualize empathy as an affective-motivational resource, moral disengagement as a cognitive mechanism, and injunctive anti-hate speech norms as social-regulative cues signaling expected and sanctioned conduct. Because these processes are embedded in classroom contexts, they may operate at both the student and classroom levels. Classroom-average empathy and moral disengagement reflect shared affective and moral orientations, whereas aggregated injunctive norms capture a collective perception that hate speech is disapproved of and countering it is supported. Importantly, classroom norms may shape how individual dispositions translate into bystander behavior. Specifically, classroom-level injunctive anti-hate speech norms are expected to moderate the association between student-level empathy and profile membership by clarifying expectations and reducing perceived social risks. From a bystander-intervention perspective, emotional and moral responses represent only part of the decision-making process, as intervention also depends on factors such as perceived responsibility, self-efficacy, and anticipated social consequences (Jost et al., 2020; Obermaier, 2024). While these mechanisms

are not measured in the present study, they highlight that translating empathy into bystander behavior is contingent on contextual and normative factors.

1.2.1 Student-Level social-cognitive and normative predictors

Empathy, defined as the capacity to understand and share others' emotional states (Decety and Cowell, 2015), was consistently associated with prosocial behavior among youth (Qiu et al., 2024). In hate speech contexts, empathy for victims may motivate responses such as comforting the victim, seeking support, or actively countering hateful remarks. Variable-centered research indicated that higher empathy is associated with more favorable bystander reactions and fewer disengaged or harmful responses (Wachs et al., 2024a). Similar associations have been observed in digital contexts, where empathy predicted adolescents' intentions to engage in online counter speech (Sportelli et al., 2025). The relevance of empathy as a predictor of bystander roles is further supported by longitudinal research showing that broader social and emotional competencies, including emotion regulation and responsible decision-making, were associated with lower involvement in online hate speech perpetration and victimization over time (Llorent et al., 2025). At the same time, emerging evidence suggests that empathy does not uniformly translate into prosocial action but may, more broadly, increase emotional engagement in bystander situations (Breithaupt, 2018; Guthridge et al., 2020), potentially leading to different forms of active responding depending on regulatory processes and contextual cues.

Moral disengagement refers to cognitive mechanisms that enable individuals to justify, minimize, or normalize harmful behavior, thereby weakening internal moral self-sanctions (Bandura et al., 1996). In hate speech situations, such mechanisms may involve blaming the victim, trivializing verbal harm, or diffusing responsibility among bystanders. Adolescents who morally disengage may therefore be less likely to perceive hate speech as requiring intervention and more likely to tolerate or legitimize it. Consistent with this reasoning, moral disengagement has been found to be positively associated with engagement in and tolerance of hate-related behaviors and negatively associated with prosocial bystander responses (Ieracitano et al., 2024; Thornberg et al., 2025; Wachs et al., 2022a, 2024a).

Beyond social-cognitive dispositions, adolescents' responses are also shaped by perceived injunctive anti-hate speech norms. Injunctive norms generally reflect individuals' perceptions of approval or disapproval from significant others (Cialdini and Trost, 1998), such as peers, teachers, or family members. In the context of hate speech, these norms capture beliefs about whether hate speech is socially sanctioned and whether defending targeted peers is supported. Although these sources can exert distinct influences, adolescents' perceptions of social approval often coalesce into a more generalized sense of what is considered acceptable within their immediate social environment. In the present study, aggregated injunctive norms are conceptualized as reflecting a shared classroom-level approval climate regarding hate speech. While this approach does not differentiate between specific sources of influence, it captures the overall normative signal perceived by students

within their classroom context. Research found that stronger perceived injunctive anti-hate speech norms are associated with lower hate speech perpetration, weaker links between witnessing and perpetrating hate speech, and greater intentions to counter harmful behavior (Piccoli et al., 2026; Wachs et al., 2022b, c). Qualitative evidence further highlighted the centrality of peer approval and conformity pressures in shaping adolescents' responses (Ballaschk et al., 2021).

1.2.2 Classroom-Level affective, moral, and normative climates

Beyond individual characteristics, bystander responses are embedded in classroom contexts characterized by shared affective, moral, and normative climates. In educational systems with stable class groupings, such classroom characteristics are reinforced through repeated interactions and may exert substantial influence on student behavior.

Classroom-level empathy and moral disengagement reflect shared affective and moral orientations. Classroom-level empathy captures collectively expressed concern for others, whereas moral disengagement reflects shared tendencies to justify or downplay harm. Research on bullying indicated that collective moral disengagement is positively associated with aggressive behavior and passive bystanding and negatively associated with defending (Gini et al., 2014, 2020; Thornberg, 2023). Hate speech research further showed that defending profiles were more prevalent in classrooms characterized by cohesion and collective intervention, whereas reinforcing or retaliatory profiles were more common in classrooms with higher rates of hate speech perpetration (Wachs et al., 2024b).

Classroom-level injunctive anti-hate speech norms represent a shared regulatory climate that defines socially expected and sanctioned behavior. At the classroom level, these norms capture collective disapproval of hate speech and shared support for intervention. Conceptually, injunctive norms differ from affective and moral climates: whereas shared empathy and moral disengagement shape emotional and moral appraisals of incidents, injunctive norms reduce behavioral ambiguity, lower perceived social risks, and provide collective support for norm-consistent action. Research in bullying and peer aggression contexts showed that classroom norms fostered defending behavior (Kollerová et al., 2018; Lucas-Molina et al., 2018). More broadly, multilevel research has demonstrated that classroom norms and moral climates shape bystander behavior and moderate associations between individual characteristics and behavioral responses (e.g., Gini et al., 2015; Lucas-Molina et al., 2018; Sentse et al., 2015). However, these processes have rarely been examined in the context of bias-based aggression, such as racist hate speech, and particularly not in relation to differentiated bystander response profiles.

1.2.3 Cross-Level dynamics: classroom norms and individual dispositions

Social Cognitive Theory emphasizes that behavior emerges from dynamic interactions between individual dispositions and social contexts (Bandura, 1986). Prior multilevel research in bullying contexts has shown that classroom norms can moderate

the association between individual dispositions, such as empathy, and bystander intervention (e.g., Lucas-Molina et al., 2018; Sentse et al., 2015). Building on this framework, the present study examined whether classroom-level injunctive anti-hate speech norms can shape the association between student-level empathy and adolescents' latent bystander response profiles.

Empathy constitutes an intrapersonal affective resource that heightens students' sensitivity to the suffering of victimized peers and fosters prosocial motivation to intervene (Qiu et al., 2024). However, whether empathy translates into observable defending behavior depends on the normative classroom context. For example, multilevel research in bullying contexts showed that classroom climate can shape the link between empathy and bystander responses. In one study, a stronger bullying climate attenuated the association between empathy and perceived responsibility to intervene (Zhang et al., 2025). In contrast, another study found that empathy predicted defending more strongly in classrooms characterized by higher levels of bullying (Peets et al., 2015), suggesting that contextual factors may influence whether empathic concern translates into action.

At the same time, strong classroom norms may be associated with lower prevalence of antisocial role patterns. When hate speech is collectively disapproved of, and intervention is socially supported, revenger and contributor roles may become less normative and less prevalent. In such contexts, individual differences in empathy may play a smaller role in distinguishing between membership in these antisocial profiles. In contrast, in classrooms with weak or ambiguous norms, antisocial responses may be more socially tolerated, increasing the extent to which low empathy is associated with revenger or contributor roles.

More broadly, these patterns are consistent with research showing that classroom norms can moderate the association between individual characteristics and bystander behavior in bullying and peer aggression contexts (Lucas-Molina et al., 2018; Sentse et al., 2015; Wang et al., 2024). Examining these cross-level interactions is therefore essential for understanding how individual affective dispositions and shared classroom norms jointly shape heterogeneous bystander responses to racist hate speech.

1.3 The present study

Building on Social Cognitive Theory (Bandura, 1986) and prior research on bystander responses to racist hate speech in school contexts (Wachs et al., 2024b), the present study examined how adolescents' heterogeneous bystander responses emerge from the interplay between individual social-cognitive dispositions and shared classroom environments. Specifically, we investigated empathy and moral disengagement as key affective and cognitive mechanisms at both the individual (student-level) and aggregated (classroom-level). In addition, we examined injunctive anti-hate speech norms, which capture perceived social expectations regarding acceptable conduct at the individual and classroom levels. Rather than focusing on isolated behaviors, the study draws on previously identified latent bystander response profiles—defenders, passive bystanders, revengers, and contributors—that reflect systematic configurations of co-occurring reactions to racist hate speech.

Although prior variable-centered research demonstrated that empathy, moral disengagement, and injunctive anti-hate speech norms are associated with specific bystander behaviors, these social-cognitive and normative factors have not yet been systematically examined in relation to differentiated bystander profiles in racist hate speech contexts. In particular, little is known about how individual dispositions and shared classroom climates jointly relate to profile membership, nor whether classroom-level injunctive norms condition the translation of individual empathy into distinct response patterns. Addressing these gaps is essential for determining whether heterogeneous bystander profiles reflect stable individual differences or contextually structured patterns of moral and normative regulation.

By integrating student- and classroom-level predictors and testing cross-level interactions within a unified multilevel framework, this study examines how heterogeneous bystander response profiles toward racist hate speech are shaped in school settings. This approach provides a theoretically grounded basis for understanding how individual dispositions and shared norms jointly influence adolescents' responses. Against this background, the following hypotheses were tested:

Hypothesis 1: Empathy, at both the student and classroom levels, is positively associated with membership in the defender profile and negatively associated with membership in the revenger and contributor profiles (relative to passive bystanders).

Hypothesis 2: Moral disengagement, at both the student and classroom levels, is negatively associated with membership in the defender profile and positively associated with membership in the revenger and contributor profiles (relative to passive bystanders).

Hypothesis 3: Injunctive anti-hate speech norms, at both the student and classroom levels, are positively associated with membership in the defender profile and negatively associated with membership in the revenger and contributor profiles (relative to passive bystanders).

Hypothesis 4: Classroom-level injunctive anti-hate speech norms moderate the association between student-level empathy and bystander role membership, such that stronger norms are expected to strengthen the positive association between empathy and defender membership and reduce the association between low empathy and revenger or contributor membership (relative to passive bystanders).

2 Methods

2.1 Sample

The sample comprised 3,225 adolescents from Germany ($n = 1,841$; 57.1%) and Switzerland ($n = 1,384$; 42.9%). Participants were enrolled in Grades 7–9, with ages approximately 12–15 years (Grade 7: 33.2%, $n = 1,070$; Grade 8: 35.6%, $n = 1,147$; Grade 9: 31.3%, $n = 1,008$). Regarding gender, 46.1% ($n = 1,487$) self-identified as boys, 51.7% ($n = 1,668$) as girls, 2.0% ($n = 64$) as gender diverse, and 0.2% ($n = 6$) did not report their gender. An immigrant background was reported by 40.3% of the participants ($n = 1,301$), whereas 59.7% ($n = 1,924$)

reported no immigrant background. With respect to family affluence, 30.8% of students ($n = 994$) reported low affluence, 35.8% ($n = 1,155$) medium affluence, and 32.4% ($n = 1,046$) high affluence. Socioeconomic status could not be determined for 0.9% of the sample ($n = 30$) due to missing data.

2.2 Measures

2.2.1 Bystander responses to racist hate speech

Bystander responses to racist hate speech were assessed using the Multidimensional Responses to Racist Hate Speech Scale (Wachs et al., 2024a), a psychometrically validated self-report instrument that measures adolescents' reactions when witnessing racist hate speech in school settings. Participants first read a vignette describing a hate speech incident (*"Please imagine the following situation: At your school, a student makes publicly insulting statements about people of a certain skin color or origin."*). They were then asked, *"What would you do in the situation described, or what have you done if you have experienced such a situation before?"* After the vignette, participants rated 21 items on a five-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). The scale comprises seven subscales reflecting distinct bystander response tendencies: comforting the victim (3 items; McDonald's $\omega = .84$), seeking help at school (3 items; $\omega = .85$), countering hate speech (4 items; $\omega = .81$), revenge (3 items; $\omega = .84$), reinforcing the perpetrator (3 items; $\omega = .65$), ignoring (2 items; Spearman-Brown $\rho = .76$), and helplessness (3 items; $\omega = .77$). These seven subscales served as indicators of latent bystander profiles in prior person-centered research (Wachs, Wettstein, et al., 2024).

2.2.2 Empathy

Empathy was assessed using a self-report scale adapted from Knauf et al. (2018), originally developed to measure affective and cognitive reactions to school bullying and cyberbullying and previously applied in the context of hate speech among adolescents (Wachs et al., 2022a). The scale's introduction was revised to explicitly refer to racist hate speech at school (e.g., *"When I see classmates being insulted or attacked by other classmates because of their skin color, origin, religion, sexual orientation, or gender ..."*). The empathy scale comprises six items that capture multiple facets of empathy, including perspective-taking, affective empathy, and empathic concern (e.g., *"I realize how badly they are doing," "It hurts me too," "It makes me want to comfort them"*). Items were rated on a five-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). Previous research with adolescents from Germany and Switzerland demonstrated excellent internal consistency for this scale (McDonald's $\omega = .91$).

2.2.3 Moral disengagement

Moral disengagement was assessed using a self-report scale adapted from Knauf et al. (2018), originally developed to measure moral disengagement in the context of school bullying and cyberbullying and previously applied to hate speech

research among adolescents (Wachs et al., 2022a). The scale introduction was adapted to explicitly refer to racist hate speech at school, using the same contextual framing as the empathy measure. The scale comprises nine items that capture key socio-cognitive mechanisms of moral disengagement described by Bandura et al. (1996), including attribution of blame, dehumanization of the victim, distortion of consequences, minimization of agency, and diffusion of responsibility (e.g., “I think it is their own fault,” “I tell myself it is probably not that bad,” “There is nothing I can do about it”). Items were rated on a five-point Likert-type scale from 1 (strongly disagree) to 5 (strongly agree), with higher scores indicating greater moral disengagement. The scale demonstrated excellent internal consistency in prior research (McDonald’s $\omega = .89$).

2.2.4 Injunctive anti-hate speech norms

Perceived injunctive anti-hate speech norms were assessed using three items administered after participants read a text-based vignette describing a hate speech scenario (Wachs et al., 2022b, c). The items measured perceived disapproval of hate speech by salient social referents and included: “My close friends do not like it when you say that about other people,” “In my family, they do not like it when you say this about other people,” and “The teachers at my school do not like it when you say that about other people.” Responses were recorded on a five-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). Internal consistency was acceptable (McDonald’s $\omega = .74$).

2.2.5 Control variables

Gender, immigrant background, and socioeconomic status (SES) were used as control variables. Gender was dichotomized (male vs. female); gender-diverse participants were excluded from the analyses due to their small number ($n = 64$). Immigrant background was assessed by asking whether the participant or at least one of their parents was born outside Germany or Switzerland. SES was measured using the Family Affluence Scale (FAS), which classifies family affluence as low, medium, or high based on household assets and family holidays (Hartley et al., 2016).

2.3 Procedure and sampling technique

After ethical approval was obtained from the University of Potsdam (Germany) Ethics Committee, data were collected between October 2020 and April 2021 using tablet-based questionnaires in Germany and online surveys in Switzerland. In Germany, the federal states of Berlin and Brandenburg constructed the sampling frame. Schools were stratified by school type [e.g., academic-track secondary schools (Gymnasium) vs. non-academic-track secondary schools (Realschule)] and then randomly selected using a probability-proportional-to-size procedure (Yates and Grundy, 1953). In Switzerland, the sampling frame was established using a contrastive sampling approach based on immigrant background (high vs. low) and geographic location (urban vs. rural).

From the resulting sampling frames, 100 schools (Germany: $n = 76$; Switzerland: $n = 24$) were randomly selected and contacted by phone and email. Recruitment ended once the predefined sampling targets were met. Overall, 40 schools agreed to participate (Germany: $n = 18$; Switzerland: $n = 22$), yielding a school-level participation rate of 40% (Germany: 24%; Switzerland: 92%).

Students in Grades 7–9 were eligible to participate. In Germany, two classes per grade were randomly selected from each participating school, whereas in Switzerland, all available classes across Grades 7–9 were invited. In addition, Swiss students enrolled in mixed-grade classrooms were invited to participate. In total, 264 classes were invited (Germany: $n = 106$; Switzerland: $n = 158$), and 236 participated (Germany: $n = 98$; Switzerland: $n = 138$), yielding an overall classroom-level response rate of 89% (Germany: 92%; Switzerland: 87%).

At the student-level, 5,836 adolescents were invited to participate (Germany: $n = 2,495$; Switzerland: $n = 3,341$), and 3,560 students completed the survey (Germany: $n = 1,841$; Switzerland: $n = 1,719$), yielding an overall response rate of 61% (Germany: 74%; Switzerland: 51%). A total of 335 Swiss students from mixed-grade classrooms across four schools were excluded from the analyses because mixed-grade classroom membership was confounded with country of origin.

2.4 Data analysis

The latent profile solution used in the present study was derived from prior analyses based on the same dataset (Wachs et al., 2024b), which identified four distinct bystander profiles (passive bystanders, defenders, revengers, and contributors). In that study, classification quality was acceptable, as indicated by an entropy value of 0.81 and high average posterior probabilities across profiles, with the lowest diagonal value being 0.88. In the present study, this established profile solution was used as a classification framework, and participants were assigned to profiles based on their most likely class membership. Profile membership was then treated as a categorical variable in subsequent analyses, allowing the examination of configurations of co-occurring bystander responses rather than isolated behaviors. Although this approach does not account for classification uncertainty, it enables the integration of profile membership into multilevel regression models with cross-level interactions. Alternative three-step approaches that adjust for classification error were not feasible within the present modeling framework; however, given the satisfactory classification quality of the original solution, this approach is considered appropriate for the current analyses.

We then conducted multilevel multinomial logistic regression models to account for the hierarchical structure of the data, with students (Level 1) nested within classrooms (Level 2). Students’ bystander response profiles (defenders, revengers, contributors, passive bystanders) were modeled as a multinomial outcome, with passive bystanders specified as the reference category. The passive bystander profile was used as the reference category because it represents the absence of active intervention and provides a theoretically meaningful baseline for comparing both prosocial and antisocial response patterns. A multilevel model was specified to simultaneously examine student-level dispositions, classroom-level characteristics, and their cross-level

interplay. At Level 1, predictors included empathy, moral disengagement, perceived injunctive anti-hate speech norms, and covariates (gender, immigrant background, and socioeconomic status). At Level 2, classroom-average empathy, classroom-average moral disengagement, and classroom-average injunctive anti-hate speech norms were included. Aggregation to the classroom level was supported by intraclass correlation coefficients, indicating meaningful between-classroom variability and acceptable reliability of classroom means (empathy: ICC1 = .11, ICC2 = .64; moral disengagement: ICC1 = .11, ICC2 = .65; injunctive norms: ICC1 = .12, ICC2 = .67).

To examine whether classroom-level injunctive anti-hate speech norms condition the association between student-level empathy and bystander profile membership, a cross-level interaction term was specified. A random slope was modeled for student-level empathy, allowing the association between empathy and profile membership to vary across classrooms. This specification reflects the theoretical assumption that the behavioral relevance of empathy varies across classrooms and is not fixed across contexts. Random slopes were not specified for other predictors because the theoretical focus of the cross-level hypothesis concerned empathy as the focal moderator-linked disposition. This random slope was regressed on classroom-level injunctive anti-hate speech norms to test cross-level moderation. [Figure 1](#) presents a schematic overview of the multilevel random-intercept and random-slope model.

Continuous student-level predictors were group-mean centered to isolate within-classroom effects, and classroom-level predictors were grand-mean centered. Main effects of student-level predictors, therefore, reflect associations at average levels of classroom characteristics. All analyses were conducted using Mplus Version 8.11 ([Muthén and Muthén, 2024](#)). Models were estimated using maximum likelihood with robust standard errors (MLR). Missing data were handled using full-information maximum likelihood. Results are reported as unstandardized logit coefficients (b), odds ratios (OR), and 95% confidence intervals.

3 Results

A multilevel multinomial logistic regression model was estimated to examine student-level and classroom-level predictors of bystander profile membership, including a cross-level interaction between student-level empathy and classroom-level injunctive anti-hate speech norms. Passive bystanders were specified as the reference category. All estimates reported below stem from this integrated model (see [Table 1](#)).

3.1 Student and classroom correlates of latent bystander profiles

At the student level, contrary to expectations, empathy was positively associated not only with defender profile membership but also with revenger and contributor membership. Specifically, empathy positively predicted defender membership ($b = 1.05$, OR = 2.85, $p < .001$), revenger membership ($b = 0.61$, OR = 1.83, $p < .001$), and contributor membership ($b = 0.32$, OR = 1.38,

$p = .008$). Higher moral disengagement was associated with a lower likelihood of defender membership ($b = -0.53$, OR = 0.59, $p < .001$) and a substantially higher likelihood of contributor membership ($b = 1.08$, OR = 2.94, $p < .001$) but was not significantly related to revenger membership ($b = 0.12$, OR = 1.12, $p = .286$). Stronger perceived injunctive anti-hate speech norms were associated with a higher likelihood of belonging to the defender profile ($b = 0.25$, OR = 1.29, $p < .001$) and a lower likelihood of belonging to the contributor profile ($b = -0.27$, OR = 0.76, $p = .003$), with no significant association for revenger membership ($b = -0.01$, OR = 0.99, $p = .889$). Gender was positively associated with defender membership ($b = 0.22$, OR = 1.24, $p = .039$) and negatively associated with revenger ($b = -1.27$, OR = 0.28, $p < .001$) and contributor membership ($b = -0.83$, OR = 0.44, $p < .001$). Immigrant background showed a negative association with defender membership ($b = -0.33$, OR = 0.72, $p = .004$) and a positive association with revenger ($b = 0.42$, OR = 1.52, $p = .001$) and contributor membership ($b = 0.47$, OR = 1.60, $p = .019$). SES showed only a positive association with defender membership ($b = 0.16$, OR = 1.17, $p = .025$).

At the classroom level, classroom-average empathy was positively associated with the likelihood of belonging to the defender profile ($b = 1.21$, $p < .001$) and the revenger profile ($b = 1.02$, $p = .001$), whereas it was not significantly associated with contributor membership ($b = 0.40$, $p = .182$). Classroom-average moral disengagement was negatively associated with defender membership ($b = -1.39$, $p < .001$) and positively associated with contributor membership ($b = 1.22$, $p < .001$) but unrelated to revenger membership ($b = 0.10$, $p = .763$). Classroom-average injunctive anti-hate speech norms were negatively associated with contributor membership ($b = -1.35$, $p < .001$), whereas no significant classroom-level main effects emerged for defender ($b = 0.13$, $p = .596$) or revenger membership ($b = -0.40$, $p = .185$).

3.2 Cross-level interaction

To examine whether classroom-level injunctive anti-hate speech norms condition the association between student-level empathy and profile membership, the random slope of empathy was regressed on classroom-level injunctive norms. Classroom-level injunctive norms significantly predicted variability in the empathy effect for defender membership ($b = -0.32$, $p = .049$). This indicates that stronger classroom-level injunctive anti-hate speech norms attenuated the positive association between individual empathy and the likelihood of belonging to the defender profile (relative to passive bystanders). Simple slope analyses showed that the empathy-defender association was strongest in classrooms with low injunctive norms ($b = 1.18$, $p < .001$), weaker at average levels ($b = 1.05$, $p < .001$), and weakest in classrooms with high injunctive norms ($b = 0.91$, $p < .001$). However, the strength of this association decreased as injunctive norms increased, suggesting that individual empathy played a less decisive role in classrooms with stronger shared anti-hate speech norms. This pattern may suggest that in classrooms with stronger injunctive anti-hate speech norms, defending behavior becomes less dependent on individual empathic disposition. As illustrated in [Figure 2](#), the positive

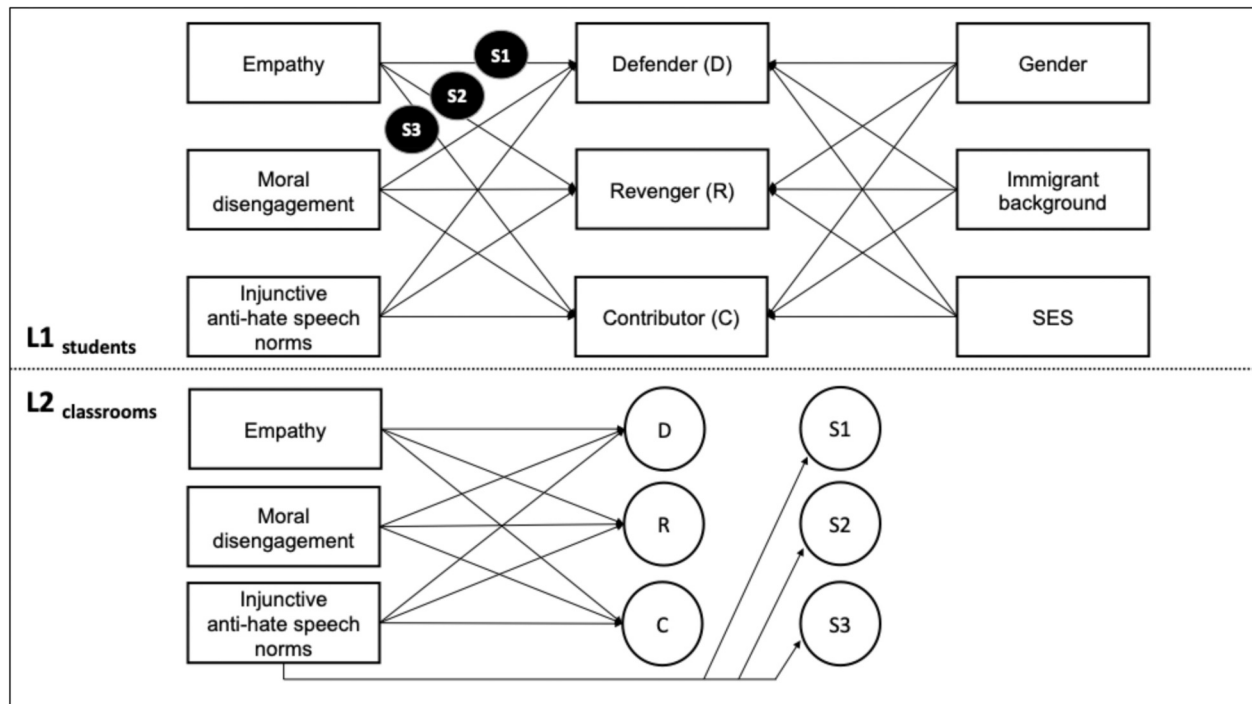


FIGURE 1
Multilevel random-intercept and random-slope model with cross-level moderation.

TABLE 1 Multilevel multinomial logistic regression predicting students' bystander profiles.

Profile Membership (Reference Category: Passive Bystander)									
Parameter	Defender			Revenger			Contributor		
	B	SE	OR [95% CI]	B	SE	OR [95% CI]	B	SE	OR [95% CI]
Individual-level									
Empathy	1.05***	0.07	2.85 [2.49, 3.26]	0.61***	0.10	1.83 [1.50, 2.24]	0.32**	0.12	1.38 [1.09, 1.74]
Moral disengagement	−0.53***	0.09	0.59 [0.50, 0.71]	0.12	0.11	1.12 [0.91, 1.39]	1.08***	0.13	2.94 [2.27, 3.81]
IAHSN	0.25***	0.06	1.29 [1.16, 1.46]	−0.01	0.09	0.99 [0.83, 1.18]	−0.27**	0.09	0.76 [0.64, 0.91]
Gender ^{girls}	0.22*	0.10	1.24 [1.01, 1.52]	−1.27***	0.17	0.28 [0.21, 0.40]	−0.83***	0.16	0.44 [0.32, 0.59]
Immigrant background ^{yes}	−0.33**	0.11	0.72 [0.58, 0.90]	0.42**	0.13	1.52 [1.18, 1.96]	0.47*	0.19	1.60 [1.08, 2.36]
SES	0.16*	0.07	1.17 [1.02, 1.36]	0.15	0.09	1.16 [0.97, 1.39]	.02	0.11	1.02 [0.82, 1.27]
Classroom-level									
Empathy	1.21***	0.21		1.02**	0.31		0.40	0.30	
Moral disengagement	−1.39***	0.24		0.10	0.32		1.22***	0.30	
IAHSN	0.13	0.24		−0.40	0.30		−1.35***	0.32	
Cross-level interaction									
Empathy x IAHSN	−0.32*	0.16		−0.26	−0.18		0.00	0.24	

IAHSN, Injunctive anti-hate speech norms, unstandardized logit coefficients (B); SE, standard errors; OR, odds ratios; CI, confidence intervals. Individual-level predictors were group-mean centered; classroom-level predictors were grand-mean centered. * $p < .05$. ** $p < .01$. *** $p < .001$.

association between empathy and the likelihood of active defending was strongest in classrooms with weaker injunctive anti-hate speech norms. Differences between simple slopes became most apparent at higher levels of empathy, reflecting attenuation of the empathy effect under stronger classroom norms rather than a qualitative change in the relationship's direction. No comparable cross-level moderation effects were observed for the revenger ($b = -0.26$, $p = .152$) or contributor profiles ($b = 0.00$, $p = .989$), suggesting that injunctive norms specifically shape the translation of empathy into norm-consistent defending rather than retaliatory or contributory responses.

4 Discussion

The present study examined how adolescents' bystander responses to racist hate speech in school settings are shaped by social-cognitive characteristics and injunctive anti-hate speech norms operating at both the student- and classroom-level. Using a multilevel, person-centered approach grounded in Social Cognitive Theory (Bandura, 1986), the study investigated associations among empathy, moral disengagement, and injunctive norms with distinct bystander response profiles, as well as cross-level dynamics linking student-level empathy with shared classroom norms. The present study provides a focused extension of prior research on bystander responses to racist hate speech by examining how individual social-cognitive dispositions relate to profile membership within classroom normative contexts. In particular, the findings offer insight into the context-dependent role of empathy and provide tentative evidence for cross-level moderation by classroom norms.

4.1 Student and classroom correlates of latent bystander profiles

Partially consistent with Hypothesis 1, empathy was positively associated not only with defender profile membership but also with revenger and contributor profiles. This finding contrasts with prior variable-centered analyses based on the same sample (Wachs et al., 2024a), which showed a negative association between empathy and antisocial responses. It also challenges the common assumption that empathy primarily promotes prosocial bystander behavior (Sportelli et al., 2025; Wachs et al., 2024a). At the classroom level, higher average empathy was likewise associated with a greater likelihood of membership in the revenger group, further underscoring the complexity of this relationship. These findings suggest that empathy may be more closely related to active engagement rather than a specific behavioral direction. Importantly, emotional and moral processes do not automatically translate into bystander behavior, as the enactment of intervention may depend on additional situational and social factors, such as perceived consequences or contextual constraints.

This pattern may also be understood in light of the person-centered approach used in the present study. Because person-centered profiles capture constellations of co-occurring responses rather than isolated behaviors, adolescents may

simultaneously express concern for the victim and endorse aggressive or reinforcing tendencies. This interpretation aligns with conceptual work on the ambivalent consequences of empathy, which suggests that empathy amplifies emotional and motivational intensity without determining its behavioral direction (Breithaupt, 2018). From a social cognitive perspective (Bandura, 1986), whether empathy translates into defending or retaliatory behavior may depend on regulatory processes and prevailing normative cues.

Despite this complexity, empathy remained a strong correlate of defender membership at both the student and classroom levels. This finding aligns with earlier work showing that empathy is associated with prosocial bystander responses, including comforting, countering, and seeking help when witnessing hate speech (Sportelli et al., 2025; Wachs et al., 2024a), as well as with longitudinal evidence linking social and emotional competencies to lower involvement in online hate speech (Llorent et al., 2025). Taken together, these results suggest that empathy plays a central role in shaping bystander behavior, but that its behavioral expression may vary depending on contextual and regulatory factors.

Largely consistent with Hypothesis 2, moral disengagement showed a predominantly maladaptive pattern across levels. At the student-level, higher moral disengagement was associated with a lower likelihood of defender membership and a higher likelihood of belonging to the contributor profile. This pattern indicates that students who more readily justify or minimize the harm of hate speech are less inclined to intervene constructively and more likely to reinforce or tolerate harmful behavior. This finding aligns closely with prior variable-centered research showing that moral disengagement undermines prosocial bystander responses while facilitating passive or reinforcing reactions to hate speech (Ieracitano et al., 2024; Thornberg et al., 2025; Wachs et al., 2022a, 2024a). In addition, classrooms with higher average levels of moral disengagement were less likely to foster defending and more likely to exhibit contributory bystander responses, largely in line with extant research on bullying and peer aggression (Gini et al., 2014, 2020; Thornberg, 2023). Conceptualized as a shared moral climate, classroom-level moral disengagement may normalize the downplaying of harm, diffusion of responsibility, or victim-blaming, thereby weakening collective self-sanctions against reinforcing or ignoring hate speech. In such contexts, disengaged interpretations of hate speech may become socially validated, reducing the likelihood that students perceive defending as necessary or expected. Together, these findings underscore that moral disengagement operates not only as an individual cognitive mechanism but also as a collective classroom characteristic that systematically constrains defending and facilitates contributory bystander behavior, consistent with Social Cognitive Theory's emphasis on moral self-regulation as a socially embedded process (Bandura, 1986).

Partially consistent with Hypothesis 3, injunctive anti-hate speech norms were positively linked to defender membership and negatively to contributor membership at the student-level. These findings indicate that students who perceived stronger normative rejection of hate speech were more likely to intervene constructively and less likely to reinforce or tolerate harmful behavior. This pattern is consistent with previous research showing that adolescents' perceived peer norms guide responses

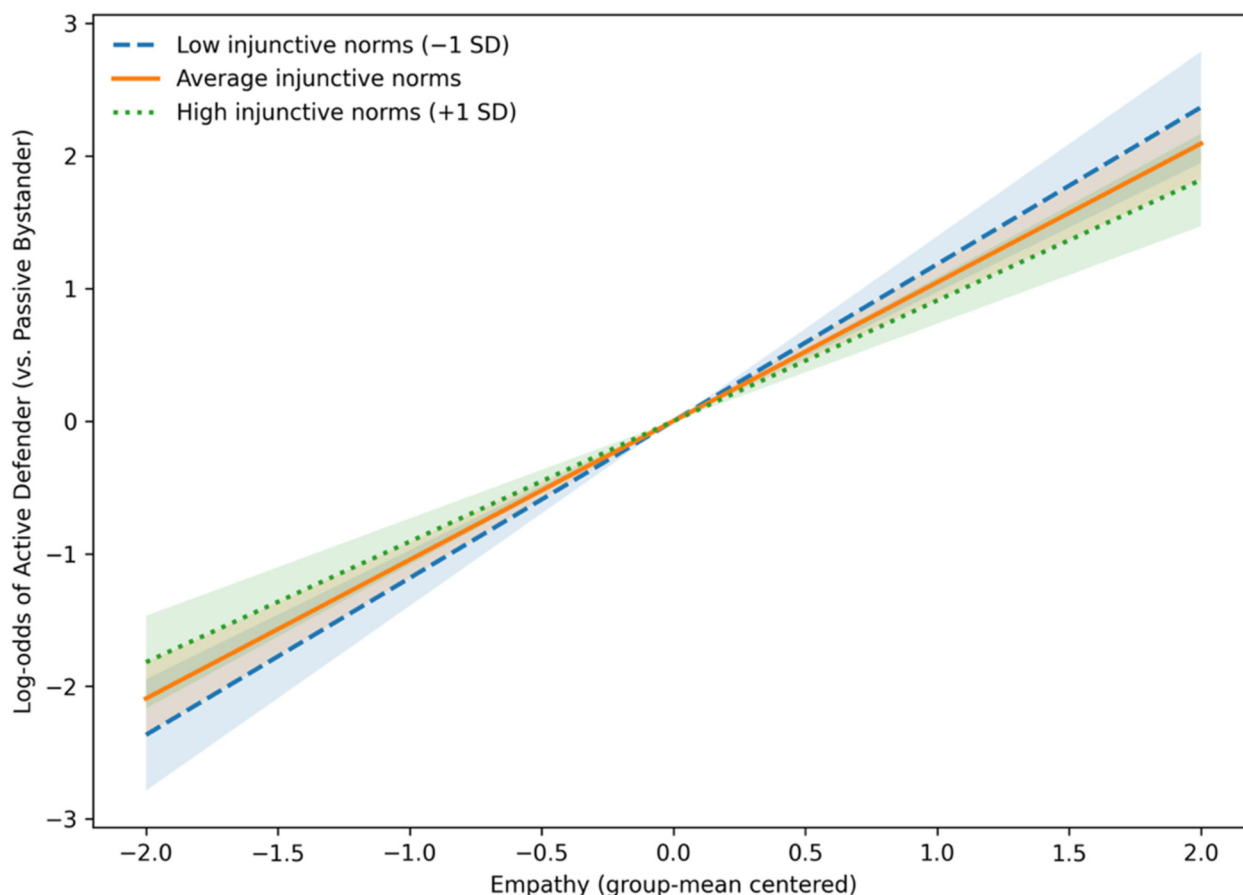


FIGURE 2

Predicted probabilities of belonging to the defender profile as a function of individual empathy and classroom-level injunctive anti-hate speech norms. The figure depicts a cross-level interaction showing that stronger classroom-level injunctive anti-hate speech norms attenuate the positive association between individual empathy and the likelihood of belonging to the defender profile (reference category: passive bystander). Lines represent model-implied predicted log-odds of defender membership as a function of student-level empathy at low (-1 SD), average, and high ($+1$ SD) levels of classroom-level injunctive norms. Shaded bands indicate 95% confidence intervals around the simple slopes.

to hate speech, reducing perpetration and increasing counter speech (Piccoli et al., 2026; Wachs et al., 2022b, c), as well as qualitative evidence highlighting the role of peer approval, status concerns, and “going along with it” dynamics in school-based hate speech (Ballaschk et al., 2021; Wachs et al., 2022b, c). This finding is also in line with research on online hate speech, which shows that spreading online hate material is positively linked to the need for online popularity, underscoring the role of norm sensitivity and anticipated peer approval in shaping hate-related behaviors (Jiménez-Díaz et al., 2025).

At the classroom-level, however, a different pattern emerged. Contrary to expectations (H3), shared injunctive anti-hate speech norms were not related to a higher likelihood of defender membership, even though they were negatively associated with contributor membership. This finding suggests that classroom-level normative disapproval of hate speech may be effective in constraining overtly harmful behavior or reinforcing bystander responses but may be insufficient on its own to promote proactive defense among students. In classrooms with strong shared disapproval of hate speech, students may be less likely to tolerate or reinforce such behavior yet still lack the confidence, skills, or collective efficacy to intervene. Thus, normative rejection may reduce contributory

responses without necessarily translating into widespread defending behavior. This finding aligns with previous research highlighting the crucial roles of social skills and self-efficacy in countering hate speech (Jost et al., 2020; Obermaier, 2024; Wachs et al., 2023).

Notably, injunctive norms were not directly associated with revenger membership at either the student- or classroom-level. This absence of association is also inconsistent with Hypothesis 3. This pattern aligns with evidence that normative beliefs against aggression predict aggressive behavior primarily when there is no provocation, whereas their regulatory influence is substantially weakened under provoking conditions (Li et al., 2015). In such situations, aggressive responding is more strongly driven by affective arousal and automatic processes than by deliberative evaluations of social approval or disapproval. This interpretation aligns with conceptualizations of reactive aggression as an affect-driven response, characterized by heightened emotional reactivity, anger, and sensitivity to perceived threat rather than reflective norm-based decision-making (Hubbard et al., 2010). In line with the previously discussed positive association between empathy and revenger membership, this pattern further suggests that emotionally engaged responses to perceived harm may facilitate retaliatory

behavior once provocation is salient. Thus, retaliatory responses may emerge even in contexts where injunctive norms clearly condemn aggression, as situational cues and emotional activation override normative restraint.

4.2 Cross-level interaction

Contrary to Hypothesis 4, classroom-level injunctive anti-hate speech norms did not strengthen the association between individual empathy and defender membership; instead, they attenuated it. This pattern suggests that in classrooms characterized by clear, strongly shared anti-hate speech norms, defending may be less contingent on individual empathic dispositions. From a Social Cognitive Theory perspective (Bandura, 1986), one possible interpretation is that strong norms reduce ambiguity and perceived social risks, which may facilitate broader defense behavior. However, given that the interaction effect was relatively small and not robust across all comparisons, this finding should be interpreted with caution and considered exploratory.

Whereas prior research has primarily conceptualized classroom norms as amplifiers of defending or as strengthening the effects of individual characteristics (e.g., Lucas-Molina et al., 2018; Sentse et al., 2015; Zhang et al., 2025), the present findings may point to a more nuanced pattern. Specifically, strong injunctive anti-hate speech norms may be associated with a reduced reliance on individual differences in empathy for predicting defending. Conversely, in classrooms with weaker or more ambiguous norms, empathy may play a more prominent role in motivating intervention. However, given the modest and not fully robust nature of the interaction effect, these patterns should be interpreted with caution.

Moreover, this attenuation effect is broadly consistent with prior cross-level findings (Peets et al., 2015), which demonstrated that emotionally intense contexts can strengthen the association between empathy and defending. Whereas heightened distress cues may amplify the motivational force of empathy, clear injunctive norms may instead structure behavior in ways that reduce reliance on individual affective dispositions. Finally, no corresponding cross-level interactions emerged for revenger or contributor profiles. This pattern suggests that injunctive norms may be more closely related to norm-consistent defending than to retaliatory or reinforcing responses, although further research is needed to clarify these associations.

4.3 Limitations

Several limitations should be considered when interpreting the current study's findings. First, the cross-sectional design precludes causal inferences about the directionality of associations among individual dispositions, classroom norms, and bystander response profiles. Although theoretical framing assumes that social-cognitive dispositions and normative climates shape bystander behavior, reciprocal processes are also plausible. For example, classrooms with a higher prevalence of defending may gradually develop stronger injunctive anti-hate speech norms. Longitudinal and experimental studies are needed to disentangle

these dynamic relationships and investigate causality. Second, the study relied exclusively on self-reported data, which may be subject to social desirability bias and shared method variance. Although self-reports are well-suited for capturing internal dispositions and perceived norms, future research could benefit from incorporating peer nominations, teacher reports, or observational data to provide a more comprehensive assessment of bystander behavior and classroom climate. Multi-informant reports could enhance the validity of the present study's findings.

Third, the focus on racist hate speech limits the generalizability of the findings to other forms of hate speech, such as sexism, homophobia, or ableism. While racist hate speech is a highly salient and socially consequential form of hate speech, future research should examine whether similar bystander profiles and normative processes emerge across different target groups. It would also be worthwhile to examine how other forms of hate speech might intersect with racist hate speech. Fourth, the measurement of injunctive norms captured adolescents' perceptions of disapproval from peers, teachers, and significant others. Although this approach reflects adolescents' lived normative environments, aggregating these perceptions to the classroom-level may not fully capture peer-specific classroom norms. Future studies should more explicitly distinguish among peer, teacher, and institutional norms and examine their potentially differential effects on bystander responses.

Fifth, a further limitation concerns the absence of measured mechanisms linking emotional and moral processes to actual bystander behavior. While the present study focused on empathy, moral disengagement, and perceived norms, it did not assess intermediate factors that may influence whether these dispositions translate into action, such as perceived social risks, responsibility, or self-efficacy. Prior research has shown that such factors play a critical role in shaping bystander intervention, with perceived responsibility and self-efficacy facilitating intervention and perceived threat inhibiting it (Jost et al., 2020; Obermaier, 2024). Future research should examine these processes more directly to better understand how bystander responses are enacted in hate speech situations. Finally, while the person-centered approach offers important advantages for capturing heterogeneity in bystander behavior, profile solutions are inherently sample-dependent. Replication in different educational settings (e.g., alternative schools), age groups, and cultural settings is necessary to establish the robustness and generalizability of the identified bystander profiles.

4.4 Practical implications

The findings of the present study have several implications for school-based prevention efforts. At the student level, empathy emerged as an important correlate of defending behavior, underscoring the relevance of fostering perspective-taking and empathy in prevention programs. At the same time, empathy was also associated with retaliatory bystander responses, indicating that emotional engagement alone does not necessarily translate into constructive intervention. This finding suggests that student-level prevention may benefit not only from strengthening empathy but also from supporting students in regulating emotional reactions and translating moral concern

into non-escalatory forms of defense. Anti-hate speech prevention programs may therefore benefit from combining empathy-focused components with training in emotion regulation, de-escalation strategies, and intervention skills.

At the classroom level, the present findings point to the importance of distinguishing among shared classroom characteristics. Injunctive anti-hate speech norms, which capture shared expectations that hate speech is disapproved of, were associated with a lower likelihood of contributory bystander responses. This suggests that clear, consistently communicated normative boundaries may help constrain behaviors that reinforce or legitimize hate speech. In contrast, classroom-level empathy and moral disengagement reflect shared affective and moral climates rather than norms *per se*. These climates may shape how hate speech incidents are interpreted and emotionally processed, but they do not prescribe behavior as injunctive norms do. Prevention efforts may therefore benefit from addressing both normative regulation (e.g., through visible teacher, parent, and peer positioning and explicit classroom agreements) and broader classroom climates that foster empathy and moral accountability.

The interplay between individual dispositions and classroom-level norms may have additional implications for inclusive prevention practices. The observed attenuation of the empathy–defending association in classrooms with stronger injunctive norms suggests that such contexts may reduce the extent to which defending depends on individual differences in empathy. This pattern is consistent with the possibility that clearly communicated and shared norms may support a wider range of students in engaging in defending behavior. However, given the modest and not fully robust nature of the interaction effects, this interpretation should be treated with caution and requires further empirical validation. Taken together, the findings suggest that anti-hate speech prevention programs may benefit from a multilevel approach that integrates student-level skill development, norm-focused classroom strategies, and efforts to cultivate supportive affective and moral classroom climates. At the same time, these implications should be interpreted cautiously, given the cross-sectional design and the modest effect sizes observed. Future research, particularly longitudinal and intervention studies, is needed to clarify how these components interact over time.

In addition, for prevention efforts to be effective, teachers need support in addressing hate speech in classroom contexts. Over the years, pre-service teacher training has been incorporated into bullying prevention efforts (e.g., Maynes and Mottonen, 2017). Teachers may benefit from training not only in identifying hate speech incidents but also in responding in ways that reinforce injunctive anti-hate speech norms and support constructive bystander responses. This aligns with recent research demonstrating that teachers' intervention strategies and self-efficacy are associated with how hate speech is addressed in school contexts (Bilz et al., 2024), highlighting the importance of strengthening educators' professional competencies as part of multilevel prevention efforts. The present findings may also inform the further development of anti-hate speech programs, such as HateLess (Wachs et al., 2025, 2026), which aims to prevent involvement in hate speech and promote defending behavior among adolescents.

5 Conclusion

The present study examined how individual dispositions and classroom-level characteristics are jointly associated with adolescents' bystander responses to racist hate speech. The findings indicate that empathy is consistently associated with active bystander involvement, but not exclusively with prosocial defending. Instead, empathy was associated with both constructive and more confrontational forms of engagement, highlighting the complexity of emotional processes in bystander situations. This pattern suggests that empathy may function as a general driver of engagement, while the specific form of bystander behavior may depend on additional regulatory and contextual factors. At the classroom level, injunctive anti-hate speech norms were associated with a lower likelihood of contributory bystander responses, suggesting the potential relevance of shared normative expectations in shaping classroom behavior. At the same time, the observed cross-level interaction between empathy and classroom norms was inconsistent with the original hypothesis and modest in magnitude, indicating that conclusions about the interplay between individual dispositions and classroom context should be drawn with caution.

Overall, the present findings provide a nuanced extension of existing research by showing that bystander responses to racist hate speech are shaped by both individual and contextual factors, but not always in straightforward or uniform ways. Rather than supporting a simple distinction between prosocial and antisocial responding, the results point to more complex patterns of engagement that may vary depending on situational and social conditions. Given the cross-sectional design and the modest effect sizes observed, the findings should be interpreted as preliminary. Future longitudinal and intervention research is needed to clarify the mechanisms linking empathy, classroom norms, and bystander behavior, and to determine how these factors can be effectively addressed in prevention efforts.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by University of Potsdam (Germany) Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

SW: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project

administration, Supervision, Validation, Writing – original draft. JK-D: Conceptualization, Data curation, Validation, Writing – review & editing. JH: Validation, Writing – review & editing. LB: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Writing – review & editing. AW: Data curation, Formal analysis, Funding acquisition, Methodology, Writing – review & editing. MG-G: Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was funded by the German Research Foundation (DFG), grant number WA 4275/2-1; BI 1046/9-1; SCHU 1370/4-1, and the University of Teacher Education Bern, grant number 19 s 0008 01.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Ballaschk, C., Wachs, S., Krause, N., Schulze-Reichert, F., Kansok-Dusche, J., Bilz, L., and Schubarth, W. (2021). „Dann machen halt alle mit.“ eine qualitative studie zu beweggründen und motiven für hatespeech unter schüler*innen [“then everyone just joins in”: a qualitative study on motives and reasons for hate speech among students]. *Diskurs Kindheits Jugendforsch./Discourse J. Child. Adolesc. Res.* 16 (4), 463–480. doi: 10.3224/diskurs.v16i4.01
- Bandura, A., and National Inst of Mental Health. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Bandura, A., Barbaranelli, C., Caprara, G. V., and Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *J. Pers. Soc. Psychol.* 71 (2), 364–374. doi: 10.1037/0022-3514.71.2.364
- Bedrosova, M., Machackova, H., Šerek, J., Smahel, D., and Blaya, C. (2022). The relation between the cyberhate and cyberbullying experiences of adolescents in the Czech republic, Poland, and Slovakia. *Comput. Human. Behav.* 126, 107013. doi: 10.1016/j.chb.2021.107013
- Bilz, L., Fischer, S. M., Kansok-Dusche, J., Wachs, S., and Wettstein, A. (2024). Teachers' intervention strategies for handling hate-speech incidents in schools. *Soc. Psychol. Educ.* 27 (5), 2701–2724. doi: 10.1007/s11218-024-09929-9
- Breithaupt, F. (2018). The bad things we do because of empathy. *Interdiscip. Sci. Rev.* 43 (2), 166–174. doi: 10.1080/03080188.2018.1450928
- Castellanos, M., Wettstein, A., Wachs, S., Kansok-Dusche, J., Ballaschk, C., Krause, N., et al. (2023). Hate speech in adolescents: a binational study on prevalence and demographic differences. *Front. Educ.* 8, Article 1076249. doi: 10.3389/feduc.2023.1076249
- Cialdini, R. B., and Trost, M. R. (1998). “Social influence: social norms, conformity, and compliance,” in *The Handbook of Social Psychology* Vol. 2, 4th ed., eds. D. T. Gilbert, S. T. Fiske, and G. Lindzey (New York, NY: Oxford University Press), 151–192.
- Decety, J., and Cowell, J. M. (2015). Empathy, justice, and moral behavior. *AJOB. Neurosci.* 6 (3), 3–14. doi: 10.1080/21507740.2015.1047055
- Gini, G., Pozzoli, T., and Bussey, K. (2015). The role of individual and collective moral disengagement in peer aggression and bystanding: a multilevel analysis. *J. Abnorm. Child. Psychol.* 43 (3), 441–452. doi: 10.1007/s10802-014-9920-7
- Gini, G., Pozzoli, T., and Hymel, S. (2014). Moral disengagement among children and youth: a meta-analytic review of links to aggressive behavior. *Aggress. Behav.* 40 (1), 56–68. doi: 10.1002/ab.21502
- Gini, G., Thornberg, R., and Pozzoli, T. (2020). Individual moral disengagement and bystander behavior in bullying: the role of moral distress and collective moral disengagement. *Psychol. Violence*. 10 (1), 38–47. doi: 10.1037/vio0000223
- Gönültaş, S., and Mulvey, K. L. (2023). Do adolescents intervene in intergroup bias-based bullying? Bystander judgments and responses to intergroup bias-based bullying of refugees. *J. Res. Adolesc.* 33 (1), 4–23. doi: 10.1111/jora.12752
- Guthridge, M., Mason, P. H., Penovic, T., and Giummarra, M. J. (2020). A critical review of interdisciplinary perspectives on the paradox of prosocial compared to antisocial manifestations of empathy. *Soc. Sci. Inf.* 59 (4), 632–653. doi: 10.1177/0539018420976946
- Hartley, J. E. K., Levin, K., and Currie, C. (2016). A new version of the HBSC family affluence scale-FAS III: scottish qualitative findings from the international FAS development study. *Child. Indic. Res.* 9 (1), 233–245. doi: 10.1007/s12187-015-9325-3
- Hay, D. F., Payne, A., and Chadwick, A. (2004). Peer relations in childhood. *J. Child Psychol. Psychiatry* 45 (1), 84–108. doi: 10.1046/j.0021-9630.2003.00308.x
- Hubbard, J. A., McAuliffe, M. D., Morrow, M. T., and Romano, L. J. (2010). Reactive and proactive aggression in childhood and adolescence: precursors, outcomes, processes, experiences, and measurement. *J. Pers.* 78 (1), 95–118. doi: 10.1111/j.1467-6494.2009.00610.x
- Ieracitano, F., Balenzano, C., Girardi, S., Gemmano, C. G., and Comunello, F. (2024). Online hate speech as a moral issue: exploring moral reasoning of young Italian users on social network sites. *Soc. Sci. Comput. Rev.* 42 (1), 25–47. doi: 10.1177/08944393231161124
- Jiménez-Díaz, O., and Del Rey, R. (2025). Cyberhate in adolescents and youths: a systematic review of labels and associated factors. *Aggress. Violent. Behav.* 81, 102023. doi: 10.1016/j.avb.2024.102023
- Jiménez-Díaz, O., Wachs, S., Del Rey, R., and Mora-Merchán, J. A. (2025). Associations between searching and sending cyberhate: the moderating role of need for online popularity and toxic online disinhibition. *Cyberpsychol. Behav. Soc. Netw.* 28 (1), 37–43. doi: 10.1089/cyber.2024.0305
- Jost, P., Ziegele, M., and Naab, T. K. (2020). Klicken oder tippen? Eine analyse verschiedener interventionsstrategien in unzielen online-diskussionen auf Facebook. *Z. Politikwiss.* 30 (2), 193–217. doi: 10.1007/s41358-020-00212-9
- Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., et al. (2023). A systematic review on hate speech among children and adolescents: definitions, prevalence, and overlap with related phenomena. *Trauma, Violence, & Abuse* 24 (4), 2598–2615. doi: 10.1177/15248380221108070

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. During the preparation of this work, the authors used ChatGPT and Grammarly for proofreading, language enhancement, and text flow improvement. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article. Further information can be provided by the first author.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Knauf, R.-K., Eschenbeck, H., and Hock, M. (2018). Bystanders of bullying: social-cognitive and affective reactions to school bullying and cyberbullying. *Cyberpsychol. J. Psychosoc. Res. Cyberspace*. 12 (4), Article 3. doi: 10.5817/CP2018-4-3
- Kollerová, L., Yanagida, T., Mazzone, A., Soukup, P., and Strohmeier, D. (2018). They think that I should defend": effects of peer and teacher injunctive norms on defending victimized classmates in early adolescents. *J. Youth. Adolesc.* 47 (11), 2424–2439. doi: 10.1007/s10964-018-0918-2
- Krause, N., Ballaschk, C., Schulze-Reichelt, F., Kansok-Dusche, J., Wachs, S., Schubarth, W., et al. (2021). „Ich lass mich da nicht klein machen!“ eine qualitative studie zur bewältigung von hatespeech durch schüler/innen. *Z. Bildungsforsch.* 11 (1), 169–185. doi: 10.1007/s35834-021-00291-w
- Lehman, B., and Zhang, C. (2025). “Bias-based bullying and hate crimes in schools,” in *The Routledge Handbook on Marginalized Groups in the United States and Their Challenges*, ed. M. Mangum (New York, NY: Routledge), 277–291.
- Li, J. B., Nie, Y. G., Boardley, I. D., Dou, K., and Situ, Q. M. (2015). When do normative beliefs about aggression predict aggressive behavior? An application of I3 theory. *Aggress. Behav.* 41 (6), 544–555. doi: 10.1002/ab.21594
- Llorent, V. J., Núñez-Flores, M., Marín-López, I., and Zych, I. (2025). Social and emotional competencies as longitudinal predictors of cyberhate in compulsory secondary school students. *Educ. Inf. Technol.* 30, 26219–26233. doi: 10.1007/s10639-025-13802-3
- Lucas-Molina, B., Giménez-Dasi, M., Fonseca-Pedrero, E., and Pérez-Albéniz, A. (2018). What makes a defender? A multilevel study of individual correlates and classroom norms in explaining defending behaviors. *School. Psych. Rev.* 47 (1), Article 0011.V47-1. doi: 10.17105/SPR-2017-0011.V47-1
- Maynes, N., and Mottonen, A. L. (2017). Bullying in schools: are pre-service teachers confident to address this? *Alberta J. Educ. Res.* 63 (4), 396–411. doi: 10.55016/ojs/ajer.v63i4.56443
- Muthén, L. K., and Muthén, B. O. (1998–2024). *Mplus User's Guide*. (8th ed.). Los Angeles: Muthén & Muthén.
- Obermaier, M. (2024). Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. *New Media Soc.* 26 (8), 4785–4807. doi: 10.1177/14614448221125417
- Pabian, S. (2025). Bystanders' perceptions on online hate speech: investigating the effects of perpetrators' justifications and the bystander's role on bystanders' attitude and prosocial intervention intentions. *Cyberpsychol. J. Psychosoc. Res. Cyberspace*. 19 (4), 1–25. doi: 10.5817/CP2025-4-6
- Palmer, S. B., and Abbott, N. (2018). Bystander responses to bias-based bullying in schools: a developmental intergroup approach. *Child. Dev. Perspect.* 12 (1), 39–44. doi: 10.1111/cdep.12253
- Peets, K., Pöyhönen, V., Juvonen, J., and Salmivalli, C. (2015). Classroom norms of bullying alter the degree to which children defend in response to their affective empathy and power. *Dev. Psychol.* 51 (7), 913–920. doi: 10.1037/a0039287
- Piccoli, V., Bianchi, M., Coladonato, R., Stragà, M., Luis, L., and Carnaghi, A. (2026). Exploring peer influence on adolescents' exposure to and engagement with online hate speech: the mediational role of perceived peers' social norms. *J. Community. Appl. Soc. Psychol.* 36 (1), Article e70223. doi: 10.1002/casp.70223
- Qiu, X., Gao, M., Zhu, H., Li, W., and Jiang, R. (2024). Theory of mind, empathy, and prosocial behavior in children and adolescents: a meta-analysis. *Curr. Psychol.* 43 (22), 19690–19707. doi: 10.1007/s12144-024-05762-7
- Rudnicki, K., Vandeboosch, H., Voué, P., and Poels, K. (2023). Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behav. Inf. Technol.* 42 (5), 527–544. doi: 10.1080/0144929X.2022.2027013
- Sentse, M., Veenstra, R., Kiuru, N., and Salmivalli, C. (2015). A longitudinal multilevel study of individual characteristics and classroom norms in explaining bullying behaviors. *J. Abnorm. Child. Psychol.* 43 (5), 943–955. doi: 10.1007/s10802-014-9949-7
- Sportelli, C., Cicirelli, P. G., Paciello, M., Corbelli, G., and D'Errico, F. (2025). Let's make the difference!" promoting hate counter-speech in adolescence through empathy and digital intergroup contact. *J. Community. Appl. Soc. Psychol.* 35 (1), Article e70028. doi: 10.1002/casp.70028
- Thornberg, R. (2023). Longitudinal link between moral disengagement and bullying among children and adolescents: a systematic review. *Eur. J. Dev. Psychol.* 20 (6), 1099–1129. doi: 10.1080/17405629.2023.2191945
- Thornberg, R., Sjögren, B., Gini, G., and Pozzoli, T. (2025). Moral disengagement and defender self-efficacy and their associations with defending, unconcerned bystanding, and guilty bystanding in school bullying. *Soc. Psychol. Educ.* 28 (1), 196. doi: 10.1007/s11218-025-10156-z
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2023). *Addressing hate speech through education: A guide for policy-makers*. Available online at: https://www.un.org/en/genocideprevention/documents/publications-and-resources/Addressing_hate_speech_through_education_A_guide_for_policy_makers.pdf (Accessed May 20, 2026).
- Wachs, S., Bilz, L., Wettstein, A., and Espelage, D. L. (2024a). Validation of the multidimensional bystander responses to racist hate speech scale and its association with empathy and moral disengagement among adolescents. *Aggress. Behav.* 50 (1), Article e22105. doi: 10.1002/ab.22105
- Wachs, S., Bilz, L., Wettstein, A., Wright, M. F., Kansok-Dusche, J., Krause, N., et al. (2022a). Associations between witnessing and perpetrating online hate speech among adolescents: testing moderation effects of moral disengagement and empathy. *Psychol. Violence*. 12 (6), 371–381. doi: 10.1037/vio0000422
- Wachs, S., Castellanos, M., Wettstein, A., Bilz, L., and Gámez-Guadix, M. (2023). Associations between classroom climate, empathy, self-efficacy, and countering hate speech among adolescents: a multilevel mediation analysis. *J. Interpers. Violence*. 38 (5–6), 5067–5091. doi: 10.1177/08862605221120905
- Wachs, S., Schittenhelm, C., Kops, M., Gámez-Guadix, M., and Wright, M. F. (2025). Happiness through HateLess? Examining the direct and indirect effects of an anti-hate speech program on victimized and non-victimized youth. *J. Adolesc.* 97 (6), 1645–1655. doi: 10.1002/jad.12525
- Wachs, S., Schittenhelm, C., Kops, M., and Wright, M. F. (2026). Promoting classroom climate and why it matters in anti-hate speech prevention programs: a HateLess evaluation study. *J. Interpers. Violence*. 41 (1–2), 93–119. doi: 10.1177/08862605241308296
- Wachs, S., Wettstein, A., Bilz, L., Espelage, D. L., Wright, M. F., and Gámez-Guadix, M. (2024b). Individual and contextual correlates of latent bystander profiles toward racist hate speech: a multilevel person-centered approach. *J. Youth. Adolesc.* 53, 1271–1286. doi: 10.1007/s10964-024-01968-x
- Wachs, S., Wettstein, A., Bilz, L., and Gámez-Guadix, M. (2022b). Adolescents' motivations to perpetrate hate speech and links with social norms. *Comunicar* 30 (71), 9–20. doi: 10.3916/C71-2022-01
- Wachs, S., Wettstein, A., Bilz, L., Krause, N., Ballaschk, C., Kansok-Dusche, J., et al. (2022c). Playing by the rules? An investigation of the relationship between social norms and adolescents' hate speech perpetration in schools. *J. Interpers. Violence*. 37 (21–22), NP21143–NP21164. doi: 10.1177/08862605211056032
- Wang, C., Li, B., Zhang, L., Liu, Y., and Xu, P. (2024). Prosocial behavior and teachers' attitudes towards bullying on peer victimization among middle school students: examining the cross-level moderating effect of classroom climate. *School. Psych. Rev.* 53 (5), 579–592. doi: 10.1080/2372966X.2021.2009313
- Yates, F., and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *J. R. Stat. Soc. Ser. A Gen.* 15 (2), 253–261. doi: 10.1111/j.2517-6161.1953.tb00140.x
- Zhang, H., Shi, C., and Dai, C. (2025). The effects of empathy and classroom norms on bystander intervention in bullying: a multilevel analysis. *Psychol. Sch.* 62 (8), 2452–2462. doi: 10.1002/pits.23479