International Journal of Developmental Science 17 (2023) 81–92
DOI 10.3233/DEV-230346
IOS Press

How to Reduce Online Hate Speech Among Adolescents? A Pilot Study on the Effects of a Teaching Unit on Social Norms, Self-efficacy and Knowledge About Hate Speech

Jan S. Pfetsch and Duygu Ulucinar

Department of Educational Psychology, Technische Universität Berlin, Germany

Abstract

A quasi-experimental study with ninth graders evaluated a 1.5-hour hate speech teaching unit in an intervention vs. control group (*N* = 82) before (T1) and after the intervention (T2). Participants reported frequency of witnessing hate speech (T1), hate speech norm and self-efficacy countering hate speech (T1 and T2), and knowledge concerning hate speech (T2). Repeated ANOVAs showed a significant three-way interaction for hate speech norm: Especially among those who witnessed hate speech more often, the program diminished the agreement to hate speech norm. Self-efficacy did not change significantly, but knowledge was slightly higher in the intervention group, particularly among students with a migration background. In sum, the intervention showed effects on norms and knowledge, and longer programs with more interactive elements for coping with hate speech seem recommendable. The current research is a first step and evidence-based practice for prevention of hate speech like the current evaluation study is desperately needed.

Keywords

Norm change, knowledge acquisition, self-efficacy, prevention evaluation study, teaching unit

Introduction

Adolescents frequently use online applications and face various risks on the Internet. According to the 4C model of online risks, children and adolescents may be exposed to different kinds of risks for their emotional, physical, or mental wellbeing when being online – content, contract, contact, and conduct risks (Livingstone & Stoilova, 2021). Content risks, like the confrontation with violent or pornographic material, misinformation or embedded marketing, received high attention from researchers and practitioners alike in former years. Nowadays, especially communication and interaction risks like online hate speech have received more concern, because of the high number of private and public communication and due to their interactive, dynamic and volatile

Author Note

We thank the school and all participants for their kind support for our study. We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to PD Dr. Jan Pfetsch, Technische Universität Berlin, Department of Educational Psychology, Fraunhoferstraße 33-36, FH 5-1, 10587 Berlin, Germany. E-mail: jan.pfetsch@tu-berlin.de

nature, which makes them difficult to control by media regulation (Brüggen et al., 2022).

Given that online hate speech is a widespread phenomenon (Hasebrink et al., 2019; Smahel et al., 2020; Wachs et al., 2019) and can elicit harm for witnesses and victims (Geschke et al., 2019; Obermaier & Schmuck, 2022; UK Safer Internet Centre, 2016), measures to reduce online hate speech seem warranted, like legal regulation, technological approaches and education (Blaya, 2019). While preventive interventions for youth are often suggested against online hate speech, they are seldom empirically tested (Blaya, 2019; Seemann-Herz et al., 2022; Windisch et al., 2022). The current study therefore empirically evaluated a preventive intervention against online hate speech for ninth graders in a German secondary school. The quasi-experimental study analyzed the impact of the 1.5-hour anti-hate speech teaching unit on hate speech norm, self-efficacy countering hate speech, and knowledge concerning hate speech.

Online Hate Speech

Online hate speech can be understood as the defamation of individuals based on assumed group membership on the internet (also called online hate, cyberhate, hate material, hate speech): "Hate speech is a derogatory expression (e.g., words, posts, text messages, images, videos) about people (directly or vicariously) on the basis of assigned group characteristics (e.g., ethnicity, nationality, gender, sexual orientation, disability, 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, societal)." (Kansok-Dusche et al., 2022, p. 11). Important characteristics of hate speech thus include the following: 1) Hate Speech as a behavior that expresses derogation about people through various media like text, images or videos, 2) It references to assigned group characteristics that might not align with the target person's self-definition, ascribed by perpetrators of hate speech based on obvious or subtle characteristics, 3) It is characterized by an intention to cause harm (which distinguishes it from discrimination).

Because online hate speech degrades individuals based on assumed group membership, it shares conceptual relations with the term group focused enmity, which resembles different outgroup-specific prejudices rooted in a general derogation of others. Prejudice against various outgroups like xenophobia, anti-Semitism, islamophobia, and the devaluation of homosexuals, disabled, or homeless persons, frequently originate from a general attitude of inequality and derogation of others and exhibit similar forms and outcomes, leading to the proposition that they constitute a syndrome of group-focused enmity (Friehs et al., 2022; Zick et al., 2008). Online hate speech and group-focused enmity are conceptually linked (Darmstadt et al., 2020; Hofmann, 2018; Wachs, Schubarth, et al., 2020), as they share the fundamental aspect of derogating others based on their group membership. However, online hate speech is communicative behavior aimed at causing harm and thus also has specific characteristics that differ from prejudices in general.

Prevalence of Online Hate Speech Among Adolescents in Germany

The prevalence of online hate speech can be described from the perspective of bystanders, perpetrators, and targets of online hate speech and we focus here on data for German youth. Concerning bystanders, 23% of the 9- to 17-year-olds (Hasebrink et al., 2019) and 26% of 12- to 16-year-olds had encountered online hate speech within the past 12 months (Smahel et al., 2020). Among adolescents aged 12 to 17 54% reported having encountered online hate speech in the past 12 months (Wachs et al., 2019). Finally, 7th to 9th graders reported witnessing online hate speech with a prevalence rate of 61.4% in the last 12 months (Castellanos et al., 2023). Although there are some commonalities of these surveys, e.g., the prevalence of witnessing online hate speech generally increased with age in adolescence, they also differed concerning the age of participants, definition, operationalization and time frame of reference to measure online hate speech. Thus, empirical studies found that German adolescents encountered online hate speech as bystanders from 23% to 61.4% in the last year.

Only few studies in Germany have assessed the prevalence of online hate speech perpetration. Roughly 11.5% of 12- to 17-year-olds admitted having perpetrated online hate speech on a single-item in the last 12 months. In a multiple regression analysis, higher age and male sex predicted more frequent online hate speech perpetration, while migration background and socioeconomic background did not (Wachs & Wright, 2019). Compared to Germany, the prevalence rate was lower among adolescents in Cyprus, Korea, and Spain, but higher in Thailand, and the USA (Wachs et al., 2019). Additionally, among 7th to 9th graders, 12.7% admitted having perpetrated online hate speech in the last 12 months (Castellanos et al., 2023). Generally, perpetration of online hate speech might be slightly higher due to social desirability effects, especially when asking youth with single-item questions. Based on the data available, about 11.5% to 12.7% of German adolescents reported perpetrating online hate speech.

Looking at the targets, 19.6% of 7th to 9th graders reported online hate speech which was targeted at themselves in the last 12 months (Castellanos et al., 2023). Other data show that about 17% of the 12- to 17-year-olds reported having been ever personally affected by online hate speech (Wachs, Gámez-Guadix, et al., 2020). Based on the same data, there was no bivariate relation of online hate speech victimization to age and sex, but migration background and lower socioeconomic status were positively related with small effect sizes (Wachs & Wright, 2019). Further research indicates that age and migration background might be relevant factors in victimization by online hate speech: A representative survey among adults in Germany found that 8% of the 18- to 95-year-olds reported having been affected ever by online hate speech (Geschke et al., 2019). In this study the prevalence was higher, the younger the participants were: 3% of 60-year-olds and older, 6% of 45- to 59-year-olds, 12% of 25- to 44-year-olds and 17% of 18- to 24-year-olds were affected by online hate speech. Also, migration background was related with the victimization by online hate speech: While 6% of participants without migration background reported being affected by online hate speech, 14% of participants with migration background were affected by online hate speech (Geschke et al., 2019). In sum, 17 to 19.6% of young persons under 18 years in Germany reported being the target of online hate speech. Because among adults the prevalence decreased the older participants were, it seems that young people are more often targets of online hate speech, possibly due to their higher media consumption. Additionally, having a migration background relates to elevated levels of online hate speech victimization.

School-based Prevention of Online Hate Speech

Not only being targeted by online hate speech can have detrimental effects, also witnessing online hate speech as a bystander can be related to various negative outcomes. These include negative emotions (Külling et al., 2021; Landesanstalt für Medien NRW, 2021), lower life satisfaction (Görzig et al., 2022), reduced trust (Näsi et al., 2015), increased prejudice through desensitization (Soral et al., 2018) and reduced prosocial behavior, e.g. donations for aid organizations (Weber et al., 2013; Ziegele et al., 2018). Given these negative outcomes for youth witnessing online hate speech, the reduction of online hate speech and the fostering of coping with online hate speech among adolescents seems recommendable (Blaya, 2019). Schools are an important place to support adolescents to reduce and cope with online hate speech, because of their educational mission supporting the personal development of young people for becoming responsible citizens in a digitalized world. Learning cognitive and social skills and building normative, democratic values is an essential part of teaching. Thus, preventive interventions for youth against hate speech in schools seem warranted, but are seldom empirically tested (Blaya, 2019; Seemann-Herz et al., 2022; Windisch et al., 2022). Indeed, a review on school-based prevention of (online) hate speech found 14 German-speaking prevention programs for school students in secondary schools, mainly in grades 8 – 10 (11–18 years) (Seemann-Herz et al., 2022). The duration of these programs was mostly between 1–4 school lessons, although some extended to 8, 12 or 18 school lessons (equivalent to 3 project days). Unfortunately, 12 out of 14 programs did not report any information about the evaluation of the program, neither concerning satisfaction or feedback from participants or implementing personnel nor a more comprehensive evaluation of mediators and outcomes of the program (e.g., online hate speech perpetration, coping skills or hate speech norms). One program had surveyed teachers and students, but did not publish the results, and the program SELMA was analyzed based on qualitative focus groups and interviews, which proved that the program fostered coping and resilience of youth (Seemann-Herz et al., 2022).

Since the publication of the review, the program “Hateless. Together against Hatred” was evaluated in a pre-post intervention-control-group-design (Wachs et al., 2023). The program consists of five modules and includes small projects on the school level against hate speech. The one-week program was shown to increase empathy with the victims of hate speech, self-efficacy toward intervening in hate speech, and counter-speech one month after the program (Wachs et al., 2023). However, this program is an exception in the actual research status. That virtually no empirical evaluations of online hate speech programs exist, is a call for action for evidence-based practice.

Components for School-based Prevention of Online Hate Speech

For the school-based prevention of online hate speech, three central components are focused: social norm concerning online hate speech, self-efficacy to deal with online hate speech, and knowledge about online hate speech.

Hate speech norm

Human behavior is inevitably social. What others expect from us, has a strong influence on our behavior, especially in contexts where no clear behavioral rules are set. Social norms can be understood as rules and standards that are accepted and expected by the members of a social group to guide behavior (Cialdini & Trost, 1998). These social norms describe either what most members of a social group do (*descriptive norm*) or what they approve or disapprove of (*injunctive norm*) (Cialdini et al., 1991). Holding an attitude toward a specific behavior generally increases the likelihood of engaging in this behavior, e.g., individuals who endorse norms about aggressive behavior are more likely to exhibit offline bullying (Cook et al., 2010) or cyberbullying (Kowalski et al., 2014).

The endorsement of social norms is related to the perpetration of (online) hate speech and to the reaction to it. For example, among adolescents who admitted to perpetrate hate speech, positive relations between social norms and motivations for perpetrating hate speech were found. In this study, the descriptive norm was predictive to more motives for perpetrating hate speech than the injunctive norm (Wachs, Wettstein, Bilz, & Gámez-Guadix, 2022). Concerning hate speech behavior itself, injunctive anti-hate speech norms were negatively related to perpetrating hate speech, and with higher values of injunctive anti-hate speech norms, the positive association between witnessing and perpetrating hate speech was weakened (Wachs, Wettstein, Bilz, Krause, et al., 2022). Further, when adults were confronted with online hate speech against working women, participants were more likely to flag a hate comment and were generally more willing to engage in counter speech if they more strongly endorsed solidarity citizenship norms (Kunst et al., 2021). Taken together, social norms are related to hate speech and counter speech behavior.

Self-efficacy countering hate speech

Self-efficacy is understood as a person's perception of being capable of successfully performing a behavior that leads to desired outcomes (Bandura, 1977). Perceiving an internal locus of control for one's own behavior even in face of obstacles or difficulties and experiencing personal efficacy is the foundation of human agency. Bandura proposed that performance accomplishments, such as engaging in new behavior in exercises, simulations and role play, can strengthen the sense of personal efficacy and are often incorporated in prevention programs. Self-efficacy is a central mediator from knowledge to action and has been studied either as a general conviction about one‘s own competencies (Jerusalem & Schwarzer, 1999) or as a domain-specific perception of being competent in certain domains. For instance, the latter includes the role of internet self-efficacy in addressing online opportunities and online risks (Livingstone & Helsper, 2010), adolescents self-efficacy for defending victims of offline bullying (Gini et al., 2022) or cyberbullying (Polanco-Levicán & Salvo-Garrido, 2021).

Interventions by adolescents and young adult bystanders against online hate speech is positively predicted by self-efficacy to intervene (Obermaier, 2022), self-efficacy for intervening in hate speech is positively related to counter speech (Wachs et al., 2023), and feeling more competent in written language (but not political self-efficacy) positively predicts the likelihood of commenting in uncivil online discussions (Jost et al., 2020). Thus, promoting self-efficacy to intervene in online hate speech seems promising for a preventing online hate speech.

Knowledge concerning hate speech

Although the common theme of online hate speech is the degradation of one or more individuals based on their assumed group membership, it manifests in various modes and forms. Online hate speech can be textual, visual or audio-visual (e.g., comments, GIFs or memes) or can take the form of negative stereotypes, dehumanization or calls for violence (Paasch-Colberg et al., 2021). Online hate speech can be either blatant or subtle, thus posing challenges in terms of identification and response. Enhanced understanding of its definitions, modes and forms can facilitate the detection and response to online hate speech through counter speech.

Among individuals aged 16 to 25, knowledge concerning hate speech predicted less destructive counter speech (but not more constructive counter speech). In this study hate speech-related knowledge was operationalized by the agreement about online hate speech, i.e. that it refers to insults against persons because they belong to a certain group (Obermaier, 2022). Generally, it is recommended to support the learning of key characteristics of online hate speech in prevention programs to reduce online hate speech and foster positive bystander intervention via counter speech (Atzmüller et al., 2019).

Aims of the Current Study

The current study aimed to empirically evaluate the effectiveness of a teaching unit in a pre-post intervention-control-group design in a school setting. The prevention measure was designed to reduce the social norm in favor of online hate speech, enhance self-efficacy in dealing with online hate speech, and increase knowledge about online hate speech. As a result, students who participated in the teaching unit (intervention group) compared to students who did not take part (control group) were expected to show. . . .

> H1: . . . a decrease in the social norm in favor for online hate speech.
>
> H2: . . . an increase in self-efficacy to deal with online hate speech.
>
> H3: . . . . more knowledge about online hate speech after the intervention.

In an exploratory way, we analyzed whether the frequency of witnessing online hate speech and migration background would moderate the effects of the intervention. This seemed plausible because students who have witnessed online hate speech more frequently already have a better understanding of the phenomenon and could be accustomed to a negative interaction tone on the internet (desensitization). Further, as online hate speech often addresses persons who are perceived as foreigners, students with migration background could be particularly affected by the topic as it is more strongly connected to their self-concept.

Method

Teaching Unit Against Online Hate Speech

Based on the theoretical approaches and empirical results described above, a prevention measure in the form of a teaching unit was developed for ninth graders of a secondary school. The second author, functioning as a student teacher, designed the 90-minute teaching unit, which was subsequently refined in collaboration with the first author. The main aims were to foster: 1) knowledge about prevalence, forms and outcomes of online hate speech, and reaction strategies against online hate speech, 2) a critical attitude towards online hate speech, and 3) the exemplary application of reactions to online hate speech to strengthen self-efficacy dealing with it.

The program included various didactic elements such as plenary discussions, small group activities, and individual work tasks. The students engaged in the following topics: First, a 4-minute video of a female TikToker discussing her experiences with online hate speech was shown and discussed in order to sensitize students to the prevalence of online hate speech in their daily lives. Second, students worked in small groups on different topics: a) the concept of online hate speech and its relation to group-focused enmity, b) forms and examples of online hate speech, c) possible outcomes of online hate speech for victims, and d) the advantages and disadvantages of action strategies in response to online hate speech (e.g., ignoring, reporting, counter speech with argumentation or humor). Third, each group presented their findings using posters developed in the group phase. All students received a worksheet with a table in which they documented important results from other groups. Fourth, students applied their acquired knowledge about action strategies by selecting memes and GIFs that resonated with them and were appropriate for specific situations. This reflection of humorous reactions (memes, GIFs) from a civil society organization's website against hate speech served to exemplify to students the appropriateness of responding to online hate speech and to increase their self-efficacy addressing online hate speech.

Sample

The study included four ninth grade school classes from an integrated secondary school (German: "Integrierte Sekundarschule") in a big German city. In a class-based approach the $N = 82$ children were assigned to the intervention or control group such that they had two school classes or $n = 41$ students each (see Table 1). Intervention and control group did not differ significantly concerning age ($t(80) = –0.332$, $p = 0.749$, $d = –0.071$ [–0.504; 0.362]), gender (Mann-Whitney-U-Test = 683.000, $p = 0.188$, $z = –1.315$), migration background ($U = 676.500$, $p = 0.068$, $z = –1.823$), language spoken at home ($U = 738.000$, $p = 0.222$, $z = –1.222$) or frequency of media use ($t(72.79) = –1.731$, $p = 0.088$, $d = –0.382$ [–0.818; 0.056]).

Procedure

The four ninth grade classes were selected based on their age group, in which online hate speech becomes more prevalent, and the possibility of including two classes into the intervention and two into the control group. All students and their legal guardians were informed about aims, data use and procedure of the study, voluntary, anonymous participation, and their right to end the participation without negative consequences (informed consent). Implementing the teaching unit within regular school classes was approved by the school administration and students could voluntarily decide to participate in the data collection or not. According to ethical guidelines and federal legislation adolescents aged 14 years and above are generally considered capable to decide themselves about their research participation. All students participated and filled in a paper-pencil-questionnaire directly before (T1) and after the teaching unit (T2). The questionnaire included socio-demographic data (T1), witnessing online hate speech (T1), online hate speech norm (T1 and T2), self-efficacy in countering online hate speech (T1 and T2), and knowledge concerning online hate speech (T2).

Instruments

The frequency of *Online Hate Speech Witnessing* in the last four weeks was measured with six items on a Likert scale from 0 *never* to 4 *nearly daily* (Pfetsch & Lietz, 2022), e.g., “I have seen pictures / memes that show that members of certain social groups are bad people.", Cronbach’s Alpha $\alpha = 0.726$, McDonald’s Omega $\omega = 0.736$.

Online Hate Speech Norm measured the approval of hate speech with six items on a Likert scale from 1 *does not apply at all* to 6 *applies completely* (adapted: Lietz et al., 2021), e.g., “People of a particular ethnic origin, religion or sexual orientation can also be ridiculed online.”, $\alpha_{T1} = 0.774$ / $\alpha_{T2} = 0.828$, $\omega_{T1} = 0.786$ / $\omega_{T2} = 0.845$.

Self-efficacy countering online hate speech assessed the belief to cope with online hate speech with six items on a Likert scale from 1 *does not apply at all* to 6 *applies completely* (adapted: Jerusalem & Schwarzer, 1999), e.g., “I am confident that I can always find a good solution for hate speech on the Internet.", $\alpha_{T1} = 0.765$ / $\alpha_{T2} = 0.881$, $\omega_{T1} = 0.763$ / $\omega_{T2} = 0.881$.

Knowledge concerning online hate speech (Ulucinar & Pfetsch, 2022) was measured via twelve multiple-choice questions concerning hate and counter speech (0 *wrong answer*, 1 *correct answer*), e.g., “How can users counter hate speech on social networks in a humorous way? A) PDF documents, B) Offensive photos, C) Memes or GIFs”, 12 questions, empirical range 5–12 (theoretical range 0–12), $M = 10.12$, $SD = 1.53$, $Md = 10$ correct answers.

Data Analysis

We computed two repeated measures ANOVAs (online hate speech norm, self-efficacy countering hate speech) and a univariate ANOVA (knowledge concerning online hate speech) to test the interaction effect of time by group with IBM SPSS 27. In an exploratory way, we analyzed whether the frequency of witnessing online hate speech or migration background as additional factors in the ANOVAs led to a significant interaction effect. *Post-hoc* statistical power were calculated with G*Power (Faul et al., 2007).

Assumptions for the statistical tests were analyzed (Field, 2013). Normal distribution was fulfilled for self-efficacy, but not for hate speech norm and knowledge. The condition of sphericity was met according to a non-significant Mauchly-test for self-efficacy and hate speech norm, variance homogeneity was given according to a non-significant Levene‘s test for self-efficacy and hate speech norm, but not for knowledge. Since group sizes were comparable across all test conditions, the repeated measures ANOVA (even with non-normality of hate speech norm) and univariate ANOVA (even with heterogeneity of variance of knowledge) were assumed to be robust (Field, 2013). In sum, assumptions were generally met, and statistical analyses could be performed without constraints.

Table 1
Descriptive Information About the Sample

Descriptive information	Intervention group	Control group
Age	*M* = 14.71, *SD* = 0.72, 14–16 years	*M* = 14.66, *SD* = 0.66, 14–16 years
Gender		
female	46%	34%
male	49%	66%
divers	0%	0%
no answer	5%	0%
Migration background		
yes	46%	27%
no	54%	73%
Language spoken at home		
German	66%	78%
German and another language	34%	22%
Media use	*M* = 3.50, *SD* = 0.67	*M* = 3.19, *SD* = 0.92

Note. Media use was measured on a scale from 0 = *never* to 6 = *several times a day.*

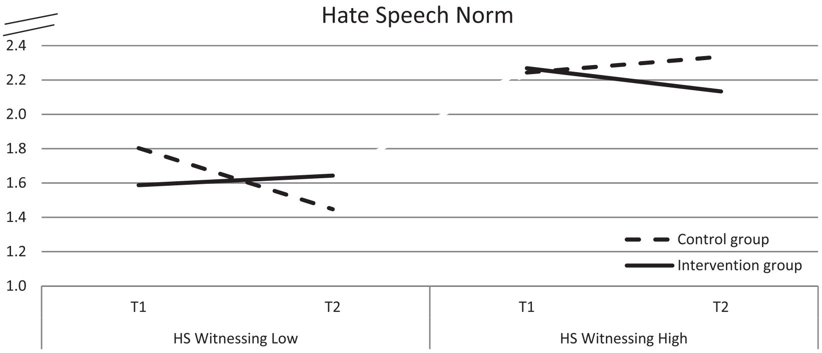
Results

Concerning hypothesis H1 about the decrease in the online hate speech norm, a repeated measures ANOVA showed a non-significant main effect for time, $F(1, 77) = 1.100$, $p = 0.298$, $\eta^2 = 0.014$ and a non-significant time × group effect, $F(1, 77) = 0.580$, $p = 0.449$, $\eta^2 = 0.007$, although test power was higher than $1 - \beta = 0.80$ (actually $1 - \beta > 0.99$). Thus, contrary to H1, the online hate speech norm did not decrease significantly in the intervention group ($M_{T1} = 1.92$, $SD_{T1} = 0.94$, $M_{T2} = 1.88$, $SD_{T2} = 0.95$) compared to the control group ($M_{T1} = 1.99$, $SD_{T1} = 0.94$, $M_{T2} = 1.83$, $SD_{T2} = 0.87$). When taking the frequency of witnessing online hate speech into account, a significant time × group × frequency of witnessing online hate speech interaction was revealed, $F(1, 76), = 4.15$, $p = 0.045$, $\eta^2 = 0.052$, see Figure 1. Especially among those students in the intervention group who witnessed online hate speech more often, the online hate speech norm decreased more strongly ($M_{T1} = 2.27$, $SD_{T1} = 0.91$, $M_{T2} = 2.13$, $SD_{T2} = 0.91$) compared to those who witnessed online hate speech less often ($M_{T1} = 1.59$, $SD_{T1} = 0.85$, $M_{T2} = 1.64$, $SD_{T2} = 0.95$).

Hypothesis H2 assumed an increase in self-efficacy to deal with online hate speech. The repeated measures ANOVA revealed a non-significant main effect of time, $F(1, 78) = 0.125$, $p = 0.725$, $\eta^2 = 0.002$, and a non-significant time × group interaction: $F(1, 78) = 0.009$, $p = 0.924$, $\eta^2 < 0.000$ (power was insufficient with $1 - \beta = 0.06$). Against the expectation, self-efficacy did not increase significantly in intervention group ($M_{T1} = 3.48$, $SD_{T1} = 0.98$, $M_{T2} = 3.48$, $SD_{T2} = 1.22$) compared to control group ($M_{T1} = 3.59$, $SD_{T1} = 0.93$, $M_{T2} = 3.60$, $SD_{T2} = 1.05$). Addition-

Figure 1

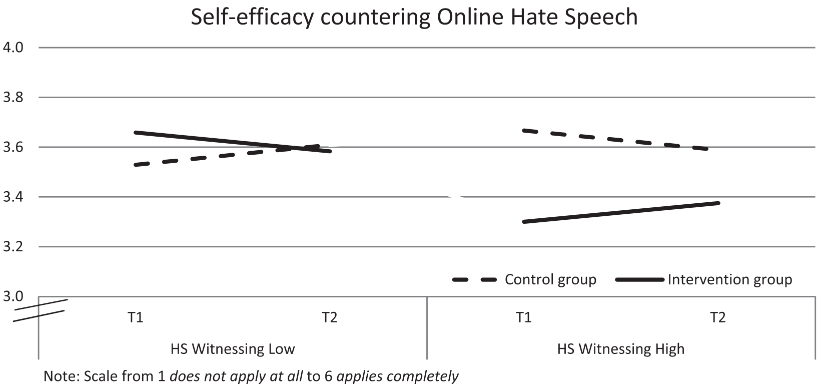
Online Hate Speech Norm in Both Groups Depending on the Frequency of Witnessing Online Hate Speech



Note: Scale from 1 *does not apply at all* to 6 *applies completely*

Figure 2

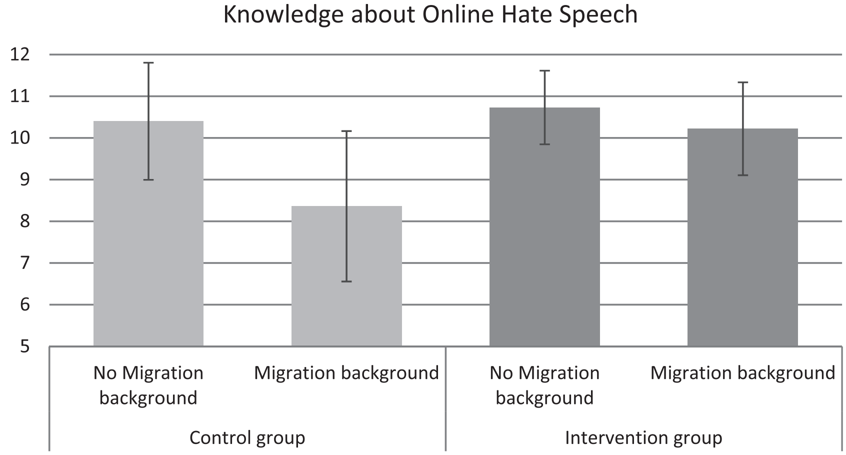
Self-efficacy Countering Online Hate Speech in Both Groups Depending on the Frequency of Witnessing Online Hate Speech



Note: Scale from 1 *does not apply at all* to 6 *applies completely*

Figure 3

Knowledge About Online Hate Speech in Both Groups Separated by Migration Background



Note: Range from 5 *only five correct answers to* 12 *all answers correct*

ally, the interaction of time × group × frequency of hate speech was not significant, $F(1, 77) = 0.776$, $p = 0.381$, $\eta^2 = 0.010$, see Figure 2.

Finally, hypothesis H3 assumed that the intervention group showed more knowledge about online hate speech compared to the control group after the prevention measure. A univariate ANOVA showed a significant effect of group, $F(1, 79) = 4.10$, $p = 0.046$, $\eta^2 = 0.049$, with a sufficient power ($1 - \beta = 0.82$). Knowledge was slightly higher in the intervention group ($M = 10.50$, $SD = 1.01$) compared to the control group ($M = 9.85$, $SD = 1.75$). Further, no effect of frequency of witnessing online hate speech was found (group × frequency), $F(1, 73) = 0.00$, $p = 0.999$, $\eta^2 < 0.000$. However, a significant group × migration background interaction was found, $F(1, 73) = 5.12$, $p = 0.027$, $\eta^2 = 0.066$, see Figure 3. Knowledge was higher for students with a migration background in the intervention group ($M = 10.22$, $SD = 1.11$) compared to students in the control group ($M = 8.36$, $SD = 1.80$).

Discussion

Online hate speech is a considerable online risk among youth. The current study tested the effect of a preventive teaching unit and found positive results for hate speech norm and knowledge concerning hate speech in subgroups. Regarding hate speech norm, a significant result was found among those who witnessed hate speech more often: The agreement to hate speech norm was significantly reduced after the teaching unit among students with more experiences regarding online hate speech, with a small effect size of $\eta^2 = 0.052$ ($0.01 < \eta^2 < 0.06$ small, $0.06 < \eta^2 < 0.14$ medium, $\eta^2 > 0.14$ large effect size; Cohen, 1988). Interestingly, those students with a higher frequency of witnessing hate speech started at a higher level of the hate speech norm, which could be due to a desensitization effect (witnessing more hate speech, agreeing more to a hate speech norm). However, participating at the teaching unit led to a decrease of hate speech norm. This differential effect indicates that the program works especially well among those ninth graders who have more experiences with online hate speech.

In contrast, hate speech self-efficacy did not change significantly through the program nor was there a significant effect when the groups were divided by frequency of hate speech ($\eta^2 = 0.01$). The lack of change in self-efficacy could potentially be attributed to the challenge of effectively altering self-efficacy within the constraints of such a short program. While the teaching unit addressed ways to react to online hate speech and included demonstrations of various forms of counter speech, the inclusion of additional practical exercises appears to be necessary to effectively enhance self-efficacy in responding to online hate speech. Given the successful augmentation of self-efficacy through programs like SELMA or HateLess, extending the intervention duration of the teaching unit seems beneficial.

Concerning hate speech knowledge, the intervention group showed higher mean scores with a small effect size of $\eta^2 = 0.049$ and smaller variance. This result was particularly pronounced for students with migration backgrounds, who experienced even greater gains in their knowledge of online hate speech, with a medium effect size of $\eta^2 = 0.066$. Given that the empirical mean scores for knowledge in the subgroups ranged from 8.4 to 10.7 (with a theoretical range from 0 *no correct answer* to 12 *all answers correct*), the items were rather easy. Nevertheless, the program's approach of cooperative group work and subsequent presentation of the results to peers contributed to the enhancement of knowledge about hate speech.

In sum, the teaching unit yielded small to medium effect sizes for main and subgroup effects for knowledge and online hate speech norm. In the related realm of cyberbullying, prevention programs significantly reduce cyberbullying victimization and perpetration with small to medium effect sizes (Hajnal, 2021; Polanin et al., 2021), but experiments to influence bystander behavior do not produce significant effects overall (Torgal et al., 2023). Adolescents feel less responsibility to intervene as online bystanders the older they get and victims who ask for help are seen as having low power, competence and social status; both aspects contribute to the fact that it is not easy to engage adolescents in prosocial bystander behaviour (Atzmüller et al., 2019). Given that fostering counter speech appears to be challenging and considering the limited duration of 1.5 hours of the teaching unit, the obtained results in the current study are still promising. However, it is clear that such a teaching unit cannot replace a comprehensive prevention program and effectively reducing online hate speech necessitates more time and a greater incorporation of interactive elements. Nevertheless, the empirical analyses of the teaching unit's effectiveness provide a basis for optimization based on these findings.

The current study has several limitations: First, the study comprised an empirical test of a teaching unit and only one main effect could be supported with a small effect size on knowledge. However, differential effects in subgroups exhibiting small to medium effect sizes could be confirmed for hate speech norm and knowledge, leading to the question how to strengthen the effectiveness of the prevention measure for all students across all outcomes. Continued engagement with the topics of hate and counter speech, coupled with more practical exercises, appears advisable. Second, the study was conducted with a limited sample size of four school classes from one age group. Apart from self-efficacy, the statistical power was sufficient for detecting the observed effects. Therefore, it is necessary to replicate these results in other age groups to ascertain their applicability in younger or older students. Third, the test items measuring hate speech knowledge exhibited a relatively low level of difficulty. This can make it more difficult to find effects due to a potential ceiling effect. While the reduced variance of knowledge in the intervention group may suggest that the students developed a common understanding of the phenomenon, employing a test with more difficult items could facilitate a clearer differentiation of knowledge levels among students.

Concerning practical implications, the prevention measure exhibited small to medium effects on the norm (especially for those with more experiences with hate speech) and knowledge about hate speech (particularly among students with a migration background). As such, this could serve as a starting point for working on the topic for teachers of ninth graders. Nevertheless, longer interventions incorporating more interactive elements, particularly those addressing coping with hate speech and developing counter speech skills, seem recommendable. Evidence-based practice for prevention and intervention of hate speech is needed to tackle the negative online risk of hate speech.

References

Atzmüller, C., Zartler, U., & Kromer, I. (2019). Online Held*innen gibt es nicht? Was 14- bis 19-jährige Jugendliche an Zivilcourage im Internet hindert [Online heroines don't exist? What prevents 14- to 19-year-olds from showing civil courage on the Internet?]. *SWS-Rundschau*, *59*, 87–109.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Blaya, C. (2019). Cyberhate: A review and content analysis of intervention strategies. *Aggression and Violent Behavior*, *45*, 163–172. https://doi.org/10.1016/j.avb.2018.05.006

Brüggen, N., Dreyer, S., Gebel, C., Lauber, A., Materna, G., Müller, R., Schober, M., & Stecher, S. (2022). *Gefährdungsatlas. Digitales Aufwachsen. Vom Kind aus denken. Zukunftssicher handeln* [Threats atlas. Growing up digitally. Thinking from the child's perspective. Acting future-proof] (2. akt. und erw. Aufl.). Bundeszentrale für Kinder- und Jugendmedienschutz. https://www.bzkj.de/resource/blob/197826/5e88ec66e545bcb196b7bf81fc6dd9e3/2-auflage-gefaehrdungsatlas-data.pdf

Castellanos, M., Wettstein, A., Wachs, S., Kansok-Dusche, J., Ballaschk, C., Krause, N., & Bilz, L. (2023). Hate speech in adolescents: A binational study on prevalence and demographic differences. *Frontiers in Education*, *8*, 1076249. https://doi.org/10.3389/feduc.2023.1076249

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. In *Advances in Experimental Social Psychology* (Vol. 24, S. 201–234). Elsevier. https://doi.org/10.1016/S0065-2601(08)60330-5

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D.T. Gilbert, S.T. Fiske & G. Lindzey (Eds.), *The handbook of social psychology* (Vols. 1-2, 4th ed., pp. 151-192). McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Cook, C. R., Williams, K. R., Guerra, N. G., Kim, T. E., & Sadek, S. (2010). Predictors of bullying and victimization in childhood and adolescence: A meta-analytic investigation. *School Psychology Quarterly*, *25*(2), 65–83. https://doi.org/10.1037/a0020149

Darmstadt, A., Prinz, M., & Saal, O. (2020). *Menschenwürde online verteidigen. 33 Social Media-Tipps für die Zivilgesellschaft* [Defending human dignity online. 33 social media tips for civil society]. Amadeu Antonio Stiftung.

Faul, F., Erdfelder, E., Lang, A. -G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics. And sex and drugs and rock „n" roll* (4th edition). Sage.

Friehs, M. -T., Kotzur, P. F., Ramos, A., & Wagner, U. (2022). Group-focused enmity – conceptual, longitudinal, and cross-national perspectives based on pre-registered studies. *International Journal of Conflict and Violence (IJCV)*, *Bd. 16*, (2022). https://doi.org/10.11576/IJCV-5361

Geschke, D., Klaßen, A., Quent, M., & Richter, C. (2019). *#Hass im Netz: Der schleichende Angriff auf unsere Demokratie* [#Hate on the Net: The creeping attack on our democracy]. https://www.idz-jena.de/fileadmin/user_upload/_Hass_im_Netz_-_Der_schleichende_Angriff.pdf

Gini, G., Pozzoli, T., Angelini, F., Thornberg, R., & Demaray, M. K. (2022). Longitudinal associations of social-cognitive and moral correlates with defending in bullying. *Journal of School Psychology*, *91*, 146–159. https://doi.org/10.1016/j.jsp.2022.01.005

Görzig, A., Blaya, C., Bedrosova, M., Audrin, C., & Machackova, H. (2022). The amplification of cyberhate victimisation by discrimination and low life satisfaction: Can supportive environments mitigate the risks? *The Journal of Early Adolescence*, 027243162210788. https://doi.org/10.1177/02724316221078826

Hajnal, Á. (2021). Cyberbullying prevention: Which design features foster the effectiveness of school-based programs? A meta-analytic approach. *Intersections*, *7*(1), 40–58. https://doi.org/10.17356/ieejsp.v7i1.648

Hasebrink, U., Lampert, C., & Thiel, K. (2019). *Online-Erfahrungen von 9- bis 17-Jährigen: Ergebnisse der EU Kids Online-Befragung in Deutschland 2019* [Online experiences of 9- to 17-year-olds: Results of the EU Kids Online Survey in Germany 2019] (2. Auflage, überarb. Auflage). Hans-Bredow-Institut.

Hofmann, A. A. (2018). Hate Speech – Gruppenbezogene Menschenfeindlichkeit im Netz [Hate Speech – Group-Focused Enmity on the Net]. *Themenblätter im Unterricht Nr. 118*. Bundeszentrale für politische Bildung.

Jerusalem, M., & Schwarzer, R. (1999). *Allgemeine Selbstwirksamkeitserwartung* [General self-efficacy expectation]. In: Ralf Schwarzer & Matthias Jerusalem (Hrsg.), *Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen (S. 16-17).* Humboldt Universität zu Berlin.

Jost, P., Ziegele, M., & Naab, T. K. (2020). Klicken oder tippen? Eine Analyse verschiedener Interventionsstrategien in unzivilen Online-Diskussionen auf Facebook [Clicking or typing? An Analysis of Different Intervention Strategies in Uncivil Online Discussions on Facebook]. *Zeitschrift für Politikwissenschaft*, *30*(2), 193–217. https://doi.org/10.1007/s41358-020-00212-9

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2022). A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, Violence, & Abuse*. https://doi.org/10.1177/15248380221108070

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, *140*(4), 1073–1137. https://doi.org/10.1037/a0035618

Külling, C., Waller, G., Suter, L., Bernath, J., Willemse, I., & Süss, D. (2021). *JAMESfocus—Hassrede im Internet* [JAMESfocus – Hate speech on the Internet]. Zürcher Hochschule für Angewandte Wissenschaften.

Kunst, M., Porten-Cheé, P., Emmer, M., & Eilders, C. (2021). Do "Good Citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics*, *18*(3), 258–273. https://doi.org/10.1080/19331681.2020.1871149

Landesanstalt für Medien NRW. (2021). *Ergebnisbericht. Forsa-Befragung zu: Hate Speech 2021* [Results Report. Forsa survey on: Hate Speech 2021]. https://www.medienanstaltnrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Hass/forsa_LFMNRW_Hassrede2021_Ergebnisbericht.pdf

Lietz, K., Pfetsch, J., & Schultze-Krumbholz, A. (2021). *Online Hate Speech – Ein Instrument zur Erfassung des Online-Verhaltens gegen*über *Gesellschaftsgruppen. Unpubliziertes Instrument [Online hate speech—An instrument to measure online behavior vis-à-vis social groups. Unpublished instrument.]*. Technische Universität Berlin.

Livingstone, S., & Helsper, E. (2010). Balancing opportunities and risks in teenagers' use of the internet: The role of online skills and internet self-efficacy. *New Media & Society*, *12*(2), 309–329. https://doi.org/10.1177/1461444809342697

Livingstone, S., & Stoilova, M. (2021). The 4Cs: Classifying online risk to children. *CO:RE Short Report Series on Key Topics*. https://doi.org/10.21241/SSOAR.71817

Näsi, M., Räsänen, P., Hawdon, J., Holkeri, E., & Oksanen, A. (2015). Exposure to online hate material and social trust among Finnish youth. *Information Technology & People*, *28*(3), 607–622. https://doi.org/10.1108/ITP-09-2014-0198

Obermaier, M. (2022). Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. *New Media & Society*, 146144482211254. https://doi.org/10.1177/14614448221125417

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

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, *9*(1), 171–180. https://doi.org/10.17645/mac.v9i1.3399

Pfetsch, J., & Lietz, K. (2022). *Beobachtung von Online Hate Speech. Unpubliziertes Instrument [Online Hate Speech Witnessing. Unpublished instrument.]*. Technische Universität Berlin.

Polanco-Levicán, K., & Salvo-Garrido, S. (2021). Bystander roles in cyberbullying: A mini-review of who, how many, and why. *Frontiers in Psychology*, *12*, 676787. https://doi.org/10.3389/fpsyg.2021.676787

Polanin, J. R., Espelage, D. L., Grotpeter, J. K., Ingram, K., Michaelson, L., Spinney, E., Valido, A., Sheikh, A. E., Torgal, C., & Robinson, L. (2021). A systematic review and meta-analysis of interventions to decrease cyberbullying perpetration and victimization. *Prevention Science*. https://doi.org/10.1007/s11121-021-01259-y

Seemann-Herz, L., Kansok-Dusche, J., Dix, A., Wachs, S., Krause, N., Ballaschk, C., Schulze-Reichelt, F., & Bilz, L. (2022). Schulbezogene Programme zum Umgang mit Hatespeech – Eine kriteriengeleitete Bestandsaufnahme [School-based programs for dealing with Hatespeech - A criteria-based inventory]. *Zeitschrift f*ür *Bildungsforschung*. https://doi.org/10.1007/s35834-022-00348-4

Smahel, D., Machackova, H., Mascheroni, G., Dedkova, L., Staksrud, E., Ólafsson, K., Livingstone, S., & Hasebrink, U. (2020). *EU Kids Online 2020: Survey results from 19 countries*. EU Kids Online. https://doi.org/10.21953/lse.47fdeqj01ofo

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

Torgal, C., Espelage, D. L., Polanin, J. R., Ingram, K. M., Robinson, L. E., El Sheikh, A. J., & Valido, A. (2023). A meta-analysis of school-based cyberbullying prevention programs' impact on cyber-bystander behavior. *School Psychology Review*, *52*(2), 95–109. https://doi.org/10.1080/2372966X.2021.1913037

UK Safer Internet Centre. (2016). *Creating a better Internet for all: Young people*'*s experiences of online empowerment 1 online hate*. http://childnetsic.s3.amazonaws.com/ufiles/SID2016/Creating%20a%20Better%20Internet%20for%20All.pdf

Ulucinar, D., & Pfetsch, J. (2022). *Wissen* über *Online Hate Speech. Unpubliziertes Instrument [Knowledge about Online Hate Speech. Unpublished instrument.]*. Technische Universität Berlin.

Wachs, S., Gámez-Guadix, M., Wright, M. F., Görzig, A., & Schubarth, W. (2020). How do adolescents cope with cyberhate? Psychometric properties and socio-demographic differences of a coping with cyberhate scale. *Computers in Human Behavior*, *104*. https://doi.org/10.1016/j.chb.2019.106167

Wachs, S., Krause, N., Wright, M. F., & Gámez-Guadix, M. (2023). Effects of the prevention program "HateLess. Together against Hatred" on adolescents' empathy, self-efficacy, and countering hate speech. *Journal of Youth and Adolescence*, *52*(6), 1115–1128. https://doi.org/10.1007/s10964-023-01753-2

Wachs, S., Schubarth, W., & Bilz, L. (2020). Hate Speech als Schulproblem? Erziehungswissenschaftliche Perspektiven auf ein aktuelles Phänomen [Hate speech as a school problem? Educational science perspectives on a current phenomenon]. In *Bewegungen: Beitr*ä*ge zum 26. Kongress der Deutschen Gesellschaft f*ür *Erziehungswissenschaft* (S. 223–236). Barbara Budrich. https://www.jstor.org/stable/j.ctv10h9fjc.19

Wachs, S., Wettstein, A., Bilz, L., & Gámez-Guadix, M. (2022). Adolescents' motivations to perpetrate hate speech and links with social norms. *Comunicar*, *30*(71), 9–20. https://doi.org/10.3916/C71-2022-01

Wachs, S., Wettstein, A., Bilz, L., Krause, N., Ballaschk, C., Kansok-Dusche, J., & Wright, M. F. (2022). Playing by the rules? An investigation of the relationship between social norms and adolescents' hate speech perpetration in schools. *Journal of Interpersonal Violence*, *37*(21–22), NP21143–NP21164. https://doi.org/10.1177/08862605211056032

Wachs, S., & Wright, M. F. (2019). The moderation of online disinhibition and sex on the relationship between online hate victimization and perpetration. *Cyberpsychol-*

ogy, Behavior, and Social Networking, *22*(5), 300–306. https://doi.org/10.1089/cyber.2018.0551

Wachs, S., Wright, M. F., Sittichai, R., Singh, R., Biswal, R., Kim, E., Yang, S., Gámez-Guadix, M., Almendros, C., Flora, K., Daskalou, V., & Maziridou, E. (2019). Associations between witnessing and perpetrating online hate in eight countries: The buffering effects of problem-focused coping. *International Journal of Environmental Research and Public Health*, *16*(20). https://doi.org/10.3390/ijerph16203992

Weber, M., Ziegele, M., & Schnauber, A. (2013). Blaming the victim: The effects of extraversion and information disclosure on guilt attributions in cyberbullying. *Cyberpsychology, Behavior, and Social Networking*, *16*(4), 254–259. https://doi.org/10.1089/cyber.2012.0328

Windisch, S., Wiedlitzka, S., Olaghere, A., & Jenaway, E. (2022). Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews*, *18*(2). https://doi.org/10.1002/cl2.1243

Zick, A., Wolf, C., Küpper, B., Davidov, E., Schmidt, P., & Heitmeyer, W. (2008). The syndrome of group-focused enmity: The interrelation of prejudices tested with multiple cross-sectional and panel data. *Journal of Social Issues*, *64*(2), 363–383. https://doi.org/10.1111/j.1540-4560.2008.00566.x

Ziegele, M., Koehler, C., & Weber, M. (2018). Socially destructive? Effects of negative and hateful user comments on readers' donation behavior toward refugees and homeless persons. *Journal of Broadcasting* & *Electronic Media*, *62*(4), 636–653. https://doi.org/10.1080/08838151.2018.1532430

Bio Sketches

Jan Pfetsch, PD Dr., Senior researcher, Department of Educational Psychology, Technische Universität Berlin, Germany. Research on offline and online bullying (esp. bystanders), learning with digital media, teacher training, vocational interests in STEM fields, empathy and prosocial behaviour in childhood and adolescence. Current research projects on digital risks for children and adolescents, technology enhanced learning in higher education, and bullying prevention.

Duygu Ulucinar, M.Ed., Teacher for Prevocational Education / Work Studies (Arbeitslehre), Department of Educational Psychology, Technische Universität Berlin, Germany. Research for her master thesis on the prevention of online hate speech.