HATE SPEECH: BASICS OF MEASUREMENT IN PRACTICE, RELATED CONSTRUCTS AND GLOBAL CITIZENSHIP EDUCATION AS A COUNTER STRATEGY

DISCORSO D'ODIO: BASI DELLA MISURAZIONE SUL CAMPO, COSTRUTTI CORRELATI E EDUCAZIONE ALLE COMPETENZE DI CITTADINANZA GLOBALE COME STRATEGIA DI CONTRASTO

Francesco Maria Melchiori
Università Niccolo' Cusano – Dipartimento di Scienze della Formazione
francesco.melchiori@unicusano.it
https://orcid.org/0000-0002-5266-7443
Sara Martucci
Laboratorio di metodologia della ricerca e analisi dei dati per le scienze del comportamento
sara.26febbraio@gmail.com

Double Blind Peer Review

Citazione

Melchiori F.M., Martucci S., (2023) Hate speech: basics of measurement in pratice, related constructs and global citizenship education as a counter strategy, Giornale Italiano di Educazione alla Salute, Sport e Didattica Inclusiva - Italian Journal of Health Education, Sports and Inclusive Didactics. Anno 7, V 1. Edizioni Universitarie Romane

Doi:
https://doi.org/10.32043/gsd.v7i1.883

gsdjournal.it
ISSN: 2532-3296
ISBN: 978-88-6022-469-9

ABSTRACT

Firstly, results of an empirical study show that there is a discrete positive correlation between hate speech and the subjects' level of empathy and awareness. Secondly, the research findings were discussed in relation to a systematic review of descriptive and experimental studies regarding hate speech.
Subsequently, a counter strategy was derived by analysing to the core concepts of global citizenship education (GCED), and the characteristics of an integrated socio-psycho-educational intervention were presented to address hate speech through a multi-faceted strategy.

In primo luogo, i risultati di uno studio empirico dimostrano che esiste una discreta correlazione positiva tra i discorsi d'odio e il livello di empatia e consapevolezza dei soggetti. In secondo luogo, i risultati della ricerca sono stati discussi in relazione a una revisione sistematica degli studi descrittivi e sperimentali sull'hate speech.
Successivamente, è stata elaborata una strategia di contrasto analizzando i concetti fondamentali dell'educazione alla cittadinanza globale (GCED) e sono state presentate le caratteristiche di un intervento socio-psico-pedagogico integrato per affrontare il discorso dell'odio attraverso una strategia multiforme.

KEYWORDS
Hate speech, global citizenship education, educational community, counter-speech
Discorso d'odio, educazione alla cittadinanza globale, comunità educante, contro-discorso

Received 4/05/2023
Accepted 11/05/2023
Published 20/05/2023

Introduction

The term “hate speech”, as defined by the Council of Europe’s Committee of Ministers, covers: “all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin” (Recommendation No. 97/20, COE, 1997). Thus, the construct of hate speech is complex and multifaceted, encompassing a range of behaviors and expressions that can be motivated by prejudice and discrimination against certain groups of people. Actually, there is not a common definition of “Hate speech” within the scientific community (MASTROMATTEI, 2022) that is capable of encompassing all the complexities and peculiarities of the case, and therefore it remains at the centre of an intense legal and academic debate at the international level, which has to come to terms with the particular subjectivity of the word and its easily manipulated description. The main difficulty is to define exhaustively all the components of hate without running the risk of colliding with some of the basic principles of democracy, including human dignity and freedom of expression (HORNSBY, 2003), concepts which, moreover, vary widely in different contemporary societies. There is also the obstacle of having to include and take into account in its description the various minority groups that are discriminated against, which are very different in their characteristics and objectively too broad to be categorised in detail. In fact, the phenomenon of hate speech was born and developed towards the end of the 1980s by lawyers who identified themselves with Critical Race Theory and were committed to exposing the racism present in US society and its legal system. Racism, like other forms of discrimination, is often generated in a context of fear of the unknown and of widespread misinformation about the facts, typically based on stereotypes and prejudices (BAGNATO, 2020). When individuals choose to belong to a particular group of subjects, they usually do so according to several variables, such as shared values and goals, individual similarities and analogies. This selection of distinctive elements is achieved through the activation of cognitive processes that are thought to be responsible for structuring how the other, the self and the world function. This categorisation is recognised and referred to as social categorisation (DE CAROLI, 2016) and manifests itself innately in the individual through the simplification and ordering of the surrounding reality, based on a dense network of similarities and differences with what has been experienced and learned throughout life. It seems important to underline that today's digital platforms characteristics facilitate the creation and dissemination of hate speech. Since they allow for rapid, effective, permanent and

inexpensive dissemination of thought, they have built an open road for the publication of any kind of message, without the presence of any structure (formal or informal) capable of exercising a mediating or controlling function. Indeed, social media permits the message to be extended to a wide audience, but they also allow the acquisition and maintenance over time of that message, which remains tracked and therefore retrievable, possibly leading to a continuous harm to the victims. According to Floridi (2017), the distinction between being online or offline is no longer relevant. Instead, we should view media as an "onlife" experience. Rivoltella & Rossi (2019) support this view by suggesting that digital technologies are not just enhancements to our experiences, but rather a natural part of our existence. This paper provides a theoretical reflection on the nomological network related to the construct of hate speech (relying on the results of a research aimed to empirical modelling/measurement and a systematic review of the literature) and its implications for strategies to promote global citizenship skills to prevent and counter the spread of this phenomenon.

1. Findings of the empirical study

A preliminary study (N=72) was conducted to create a scale to assess awareness and recognition of hate speech, which was administered to N=146 with a female prevalence (77.6%) and an average age of 22.25 years. An online questionnaire was then produced and distributed mainly in university classrooms and through snowball sampling to the population. The questionnaire includes the scale directly developed for Hate Speech detection ($\alpha = .92$) and Self-awareness of social media influence (McDonald's $\omega = 0.62$) (HATE, AWA), as well as other already validated scales measuring psychological constructs (EMP, SOC_DES, S_ESTEEM), with the aim of assessing the relationship between hate speech, awareness of the issue, empathy, social desirability and self-esteem of the subjects H1. To observe empathy, the Empathic Experience Scale (Innamorati et al., 2019) was used, a psychometric scale that measures two factors of empathy called intuitive understanding and vicarious experience. Self-esteem was measured using the Italian version of Rosenberg self-esteem scale (Prezza et al. (1997). Social desirability was measured by a short form of the Marlowe and Crowne scale, and specific items were formulated to monitor awareness of the phenomenon. With regard to hate speech, a special section was created in which participants indicated their subjective perception of hate in comments previously selected using the Delphi method. A reliability analysis was conducted on the self-developed scale of

on a smaller sample (N=38) about two months later. The values of the test-retest index were low but statistically significant (r = 0.649 p < .001).

The first statistical analysis calculates the correlations between awareness of the phenomenon, recognition of hate speech, social desirability, empathy, and self-esteem. The data revealed a series of stastistically significant correlations that confirmed the hypothesized theoretical pattern: a moderate correlation between awareness of the phenomenon and recognition of hate speech (r = .336), a moderate correlation between social desirability and self-esteem (r = .321), a small correlation between empathy and awareness of the phenomenon (r = .214), and a small correlation between empathy and recognition of hate speech (r = .269).

	HATE_TOT		AWA_TOT		S_ESTEEM_TOT		EMP_TOT		SOC_DES_TOT	
HATE_TOT	—									
AWA_TOT	0.336	***	—							
S_ESTEEM_TOT	0.011		0.053		—					
EMP_TOT	0.269	**	0.214	**	-0.018		—			
SOC_DES_TOT	-0.030		0.106		0.321	***	0.118		—	
Note. * p < .05, ** p < .01, *** p < .001										

Table 1 Correlation matrix

Subsequently, a multiple linear regression was run, where HATE_TOT was the dependent variable and Gender, Age, Education Level, Hours on socials, AWA_TOT, S_ESTEEM_TOT, EMP_TOT, and SOC_DES_TOT were the independent variables or predictors. The overall model fit measures (F 8,137 = 6.10) indicate it is statistically significant at p < .001, and R value (0.513) indicates that there is a moderate correlation between the observed values of HATE_TOT and the values predicted by the model. The adjusted R^2 value of 0.220 takes into account the number of predictors in the model and is a better measure of how well the model fits the data when there are multiple predictors. The second part of the table shows the model coefficients for each predictor. Overall, this model suggests that Gender (Female versus Male), Age, Education Level, and AWA_TOT have statistically significant relationships with HATE_TOT at p < .05.

Model Fit Measures							
				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.513	0.263	0.220	6.10	8	137	< .001

Model Coefficients - HATE_TOT						
Predictor		Estimate	SE	t	p	Stand. Estimate
Intercept [a]		3.9063	3.5086	1.113	0.268	
Gender:						
Female – Male		2.2534	0.8111	2.778	0.006	0.5500
Age		-0.1568	0.0756	-2.074	0.040	-0.1837
Education Level		1.4357	0.5811	2.471	0.015	0.2143
Hours on socials		-0.2028	0.1653	-1.227	0.222	-0.0953
AWA_TOT		0.3793	0.1291	2.938	0.004	0.2283
S_ESTEEM_TOT		0.0261	0.0693	0.377	0.707	0.0297
EMP_TOT		0.0390	0.0201	1.946	0.054	0.1517
SOC_DES_TOT		-0.0686	0.0491	-1.398	0.164	-0.1156

[a] Represents reference level

Table 2. Multiple linear regression: Hate speech detection as dependent variable

2. Systematic review of hate speech interventions

Hate speech is recognised in two different conceptions, online and offline, although it is generally more likely to be found in the online environment, so the main existing interventions relate to counter and prevention projects, explaining and dictating rules for the correct use of digital platforms. We can point out that four types of strategies to prevent/combat cyberhate can be found in the literature

- Strengthening the legal framework;
- Automated identification of cyberhate in order to regulate and intervene online;
- Education for a conscious and ethical use of the Internet and/or citizenship, and education to prioritise information on the Internet;
- Counter communication (empowering young people to produce counter discourse) (Blaya, C. 2019).

This article identifies two main types of research within the scientific framework, one relating to descriptive studies that report on the main issues of hate speech and their proposed intervention to prevent and/or resolve the situation, and another that focuses on a specific methodology tested on well-defined groups of subjects. Both are summarised and categorised in two tables (Table 3; Table 4), which will be explained in detail in the following paragraphs.

3. Descriptive studies on hate speech

At present, as far as the legal framework is concerned, we can see that in Italy there are still no specific laws concerning the virtual world, unlike, for example, Germany, which has included The German Network Law, which came into force in January 2018, which did not impose any obligations on social media platforms but introduced high fines for those who do not comply with existing legal obligations. On the other hand, with regard to the European Union more generally, on 9 December 2021, the Commission published the communication 'A more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime' (European Commission, 2022), which aims to stimulate a Council decision extending the current list of so-called EU crimes as set out in Article 83 TFEU to hate speech and hate crime. Such a decision would allow the Commission, at a later stage, to strengthen the legal framework on combating hate speech and hate crimes throughout the EU. Recall also, how the European Commission launched its own Code of Conduct in May 2016 together with four major IT companies (Facebook, Microsoft, Twitter and YouTube) in an attempt to respond to the proliferation of racist and xenophobic hate speech online (Republic Senate 2022). The purpose of the Code is to ensure that content removal requests are handled quickly. When companies receive a request to remove content deemed illegal from their online platform, they assess the request against their own rules and EU guidelines and, where necessary, national laws, which transpose EU law on combating racism and xenophobia. The companies undertake to review most of these requests in less than 24 hours, and to remove the content if necessary, always respecting the fundamental principle of freedom of speech. To date, eight companies have adhered to the Code, namely Facebook, YouTube, Twitter, Microsoft, Instagram, Dailymotion, Snapchat and Jeuxvideo.com (Senato della Repubblica, 2022). It should be noted, however, that the reports come from the platform users themselves, who, in addition to using the virtual medium for their own personal purposes, should also protect and make the network environment safe, although most people do not know how to do this. In Italy, as in the rest of Europe, the data on the phenomenon of hate speech are not reassuring: xenophobia, Islamophobia, anti-Semitic and racist speech are on the rise, especially since 2016, accomplices of the serious humanitarian crisis that has hit the Old Continent and the recent terrorist attacks (Bortone & Cerquozzi, 2017). Moreover, an increasing number of public figures, such as influencers or politicians, are using digital platforms for their own personal propaganda, often spreading hate messages against opponents or other categories, as well as generating and promoting hate speech among other users. So far, the network has been seen as a

medium for the dissemination of ideas, and the direct perpetrators are often politicians themselves, while at other times it is used as a kind of discriminatory debate among other users, legitimised by freedom of expression. In reality, however, it is not a simple exchange of opinions based on validated information, but rather insults, threats, devaluations and fake news, useful to support one's own thinking and orientation. Taking into account the current situation, with reference to counter-strategies, several strategies are proposed to eliminate hate speech when it is already present (Gagliardone et al. 2015):

- Monitor hate speech on a territorial basis;
- Developing the capacity of individuals to recognise hate in its various manifestations;
- Encouraging and facilitating reporting to the relevant authorities;
- Raising awareness about platforms that host hate speech;
- Implementing educational pathways that can develop critical awareness in individuals.

In order to intervene, many organisations have proposed to eliminate and control the messages conveyed online through specific software that uses keywords, such as the UNAR (National Office for Anti-Racial Discrimination), which in November 2015 created the Media and Internet Observatory, which aims to monitor daily not only the content of the main social media (Facebook, Twitter, GooglePlus, Youtube), but also articles, blogs and forum comments that may incite hatred and intolerance. Many associations have developed campaigns and initiatives to raise awareness among internet users about combating online hate and violence, and to improve mechanisms for monitoring and reporting cases of hate speech (Bortone & Cerquozzi, 2017). Examples include the European project eMORE (monitoring and reporting on online hate speech in Europe), coordinated in Italy by the IDOS Study and Research Centre; the project BRICkS - Building Respect on the Internet by Combating Hate Speech, which aims to provide young people with the necessary tools to critically analyse the information disseminated by online media and social networks and to promote their active role in the fight against racist and xenophobic speech online; and the European campaign "Silence Hate! Changing Words changes the World", launched on 21 March 2018, which aims to draw attention to the need to prevent the spread of hate and promote a conscious use of the web. According to Bagnato (2020), knowledge of the characteristics of hate speech and its underlying factors is essential for the implementation of any educational action aimed at effectively preventing and combating it. For this reason, any strategy to combat or prevent hate speech should be preceded by awareness-raising

programmes aimed at developing a high level of knowledge of hate speech and a sense of citizenship among the target groups. This author proposes literacy pathways that are able to develop in users the ability both to access technological tools and the web in a correct way and to understand, criticise and create non-discriminatory online content. In fact, it has been observed that the work in schools is fundamental, especially for children who are affected by the influence of the environment, so much so that it is proposed to develop the knowledge of the mechanisms of operation and the critical awareness necessary to exercise full digital citizenship, through a multidisciplinary approach integrated in the training path of civic education. The action envisaged is to take young people seriously so that they themselves take seriously the consequences of their actions, as a response to the trivialisation of content and the deresponsibilisation of attitudes, thus making young people develop responsibility, remembering that nothing disappears in the network and asking them to make a cognitive and emotional effort (Marinelli, 2021). It is therefore essential to teach young people to be responsible and critical about what they write and what they decide to publish online: that is, to be fully aware of what it means to make a comment public and of the possible consequences that may follow (Bagnato, 2020).

Reference	Population & sample	Data-Analisis	Construct	Findings
Roberto Bortone, Francesca Cerquozzi (2017). L'hate speech al tempo di Internet.	Italian and European population	Analysis of norms and interventions on the ground	Hate speech	Common awareness and monitoring of hate speech
Alberto Marinelli (2021). Educare alla cittadinanza digitale nell'era della platform society.	Italy	Analysis of digital problems	Hate speech, cyberbullying and sexting	Media education, digital citizenship, social networking and responsibility education
Wachs S, Machimbarrena JM, Wright MF, Gámez-Guadix M, Yang S, Sittichai R, Singh R, Biswal R, Flora K, Daskalou V, Maziridou E, Hong JS, Krause N. (2022). Associations between Coping Strategies and Cyberhate Involvement: Evidence from Adolescents across Three World Regions.	6829 adolescents aged 12–18 years old (Mage = 14.93, SD = 1.64; girls: 50.4%, boys: 48.9%, and 0.7% did not indicate their gender) from Asia, Europe, and North America	SEM	Cyberhate, hate speech, coping strategies and counter-speech	Creation of a survey investigating the relationship between coping strategies and counter-speech affirming that it is important to address adolescents' ability to cope with cyberhate to develop more personalized prevention approaches.
Wachs, S., Castellanos, M., Wettstein, A., Bilz, L., & Gámez-Guadix, M. (2023). Associations Between Classroom Climate, Empathy, Self-Efficacy, and Countering Hate Speech Among Adolescents: A Multilevel Mediation Analysis.	3,225 students in Grades 7 to 9 (51.7% self-identified as female) from 36 schools in Germany and Switzerland.	SEM	Classroom Climate, Empathy, Self-Efficacy, and Countering Hate Speech (counter-speech)	The findings highlight the need to focus on contextual and intrapersonal factors when trying to facilitate adolescents' willingness to face hate speech with civic courage and proactively engage against it.

Table 3 - Descriptive studies on interventions to combat and resolve hate speech.

Two main descriptive studies useful for understanding the correlates of hate speech are highlighted, both led by Wachs S., but at different times and with different

collaborators, as highlighted in Explanatory Table 1. The less recent study analysed the relationship between hate speech and different coping strategies to see the propensity of German students to be victims, accomplices or perpetrators and showed that adolescents who endorsed distal advice or endorsed technical coping were less likely to be victims, accomplices or perpetrators. In contrast, when adolescents felt powerless or endorsed retaliation to cope with cyber-hate, they were more likely to be involved in cyber-hate as victims, perpetrators or victim-authors (Wachs et al. 2022). However, the findings confirm the importance of addressing adolescents' ability to cope with cyberhate in order to develop more personalised prevention approaches, focusing on education that teaches them to practice distal counselling and technical coping when they experience cyberhate, and proposing evidence-based cyberhate prevention education (e.g. online educational games, virtual learning environments). The second study also focuses on adolescents, but goes further by examining the direct and indirect links between one contextual factor (classroom climate) and two intrapersonal factors (empathy for victims of hate speech, self-efficacy to intervene in hate speech) to understand adolescents' counterdiscourse, i.e. how much hate speech is countered. It does this by using a self-report questionnaire that includes all the constructs mentioned, which was administered in schools and reports on how classroom climate, empathy for victims of hate speech and self-efficacy to intervene in hate speech have a positive effect on countering hate speech. In addition, classroom climate has been indirectly linked to countering hate speech incitement through increased empathy and self-efficacy, so much so that the authors themselves emphasise that prevention programmes should focus on contextual and intrapersonal factors when attempting to facilitate young people's willingness to confront hate speech incitement with moral courage and to engage proactively in countering it (Wachs et al. 2023).

4. Studies with interventions and their impact on hate speech

Interventions with effects include a test-retest conducted at the national level in Poland, in which pre-selected examples of hate speech from the Internet and other mass media were presented and willingness to support a ban on public expression of such speech was assessed. The two studies confirmed this positive correlation, but showed different effects on hate speech bans. Social dominance orientation was positively correlated with acceptance of hate speech, while right-wing authoritarianism was positively correlated with prohibition of hate speech. The most likely explanation is that right-wing authoritarians are particularly vigilant

against norm violations, which makes them more punitive towards unconventional expressions of prejudice, such as hate speech (Bilewicz t al. 2015). Therefore, the evidence gathered in this study suggests that, in contrast to social dominance orientation, which has a clear negative effect on intergroup relations, right-wing authoritarianism could be a double-edged sword: it increases prejudice by increasing distance from outgroup members; at the same time, however, it could improve intergroup relations by mobilising people against such forms of prejudice that transgress their norms. The same authors believe that hate speech reduction campaigns, as long as they address the normative aspect of authoritarianism, can be an effective tool in combating hate speech. This suggestion is also supported by other studies which suggest that, notwithstanding the clearly negative effects of authoritarianism on people's attitudes towards others, there are also remarkably positive effects of authoritarianism on psychological well-being, which appear to reduce psychological distress in depression and ageing (Van Hiel & De Clercq, 2009). When considering methods to raise awareness of hate speech among citizens, it seems particularly effective to use the medium and the way in which it is disseminated, through the media. There is reason to believe that a tool powerful enough to facilitate genocide has the potential to be a tool for positive change (Vollhardt et al., 2006). One particular intervention used the media to assess its impact on the population when used in a healthy way, with the primary aim not to provide trivial theoretical analysis, but rather practical knowledge that increases competence in identifying, deconstructing and countering hate incitement. It should also provide specific knowledge and media literacy for societies in conflict as analytical tools to detect and counter hate incitement in its early stages. This article describes a short-term interdisciplinary radio campaign to raise awareness of hate incitement in the Democratic Republic of Congo (DRC) and proposes a long-term, nationwide media campaign to educate citizens and warn them of the dangers of incitement to violence. It highlights the importance of providing emotional support and solidarity when communicating with members of groups targeted by hate incitement. It is important to show awareness and address the nature of the issues involved and the fact of hate incitement, even if such discussion is sensitive and delicate. In particular, it should be expressed that the nature of the accusations and derogations has been acknowledged and is not shared. Exploratory research has shown that such solidarity and support from members of non-targeted groups can mitigate the harmful effects of targeting. Moreover, when solidarity and rejection of derogations is expressed by someone who belongs to the same social group as those who use hate speech, it makes it less likely that the entire group will be perceived as antagonistic, thus reducing the potential for cycles of violence

(Vollhardt J., Coutin M. et al 2006). This is not the only study that has used the media as an intervention against hate speech; in fact, the Dutch NGO 'Stichting Radio La Benevolencija/Humanitarian Tools Foundation' (La Benevolencija 2005), led by George Weiss and in collaboration with psychologists Ervin Staub and Laurie Anne Pearlman, launched a large-scale media campaign in Rwanda in 2003. The campaign consisted of a series of reconciliation radio programmes based on an innovative combination of a healing, reconciliation and non-recurrence of violence approach developed and implemented in Rwanda by Staub and Pearlman (Straub E., Pearlman et al. 2006). After an evaluation showed measurable positive effects of the combined approach (Paluck 2006), these programmes were later extended to Burundi and the DRC.

Reference	**Population**	**hypothesis**	**Action**
Johanna Vollhardt, Marie Coutin, Ervin Staub, George Weiss, and Johan Deflander (2006). Deconstructing Hate Speech in the DRC: A Psychological Media Sensitization Campaign;	Citizens of the Democratic Republic of Congo (DRC).	1. the definition of hate speech and the markers that can be used to detect it and distinguish it from more neutral speech. 2. the role of politicians, the media, and citizens in developing and counteracting hate speech. 3. what Congolese citizens can do to resist and counteract hate speech.	Identify, deconstruct and counter hate incitement. Counter incitement to hatred in the short term during the campaign with a series of programs broadcast weekly. Long-term media campaign throughout the country to educate citizens and warn against the dangers of incitement to violence.
Michal Bilewicz, Wiktor Soral, Marta Marchlewska, Mikołaj Winiewski (2015). When Authoritarians Confront Prejudice. Differential Effects of SDO and RWA on Support for Hate-Speech Prohibition	N 5653 adolescents; N 51007 adults; in Polonia	1. Assess the differences between two main personality antecedents of hate speech prejudice: right-wing authoritarianism (RWA, Altemeyer, 1988) and social dominance orientation (SDO, Pratto et al., 1994)	Present participants with pre-selected examples of hate speech from the internet and other mass media and assess their willingness to support a ban on public expression with such characteristics.

Wachs, S., Krause, N., Wright, M.F. et al. (2023). Effects of the Prevention Program “HateLess. Together against Hatred” on Adolescents’ Empathy, Self-efficacy, and Countering Hate Speech.	820 adolescents between 12 and 16 from 11 German schools participated in this study. More specifically, 567 adolescents participated in the one-week prevention program, and 253 participants were assigned to the control group.	1. it was hypothesised that reported levels of empathy would increase in the intervention group but not in the control group. 2. reported levels of self-efficacy would increase in the intervention group but not in the control group; 3. reported levels of counter-discourse would increase in the intervention group but not in the control group; 4. being in the intervention group would positively predict higher levels of counter-discourse through empathy and self-efficacy.	Prevention program "HateLess. Together Against Hate" a multi-level program that combines individual-level, classroom-level, school-level and community-level activities.

Table 4 - Interventions made on hate speech.

As an intervention in schools, we propose a project related to the above-mentioned descriptive studies, involving 11 German schools, which aims to assess short-term effects on adolescents' empathy, self-efficacy and counter-discourse, "The HateLess. Together against Hate" (Wachs et al. 2023). This study shows the success of HateLess, as there was a significant increase in empathy, self-efficacy and counter-discourse in the intervention group from pre-test to post-test, one month after the intervention, while no changes were found in the young people in the control group. The intervention consists of five modules, each scheduled for one school day, with three components of 90 minutes each. Counter narratives are addressed by introducing the concept of moral courage, increasing participants' sense of responsibility to counter hate speech in the classroom, and reducing passive observation. According to the authors, short stories about hate incidents could also increase young people's ability to counter hate speech. In addition, HateLess includes training in non-violent communication, including reflection on identifying feelings and exercises on expressing needs without hurting others or formulating criticism without hurting others. Participants also learn to assess when a counter discourse is recommended, for example by discussing fictitious online comments to determine whether they are hate speech or expressions of opinion,

thus learning the differences between critical but legitimate expressions and hate speech (Wachs et al. 2023). Overall, the findings are also reflected in research that has found a negative association between empathy and intolerance and prejudice, and a positive association between empathy and prosocial behaviour (Boag et al. 2008). The study is also supported by other research showing a negative association between empathy and hate speech (Celuch et al. 2022) and a positive association between empathy and counter-speech (Wachs et al. 2023). In conclusion, it can be said that the development of critical thinking skills and the ethical use of social media are the main points of any intervention, precisely because awareness of the phenomenon, integrated with empathy, seems to be the most effective construct to combat hate speech and should therefore be considered as the main starting point for media and information literacy. The expectation is that these media and information literacy skills can improve the ability of individuals to identify and challenge content that incites online hate, to understand some of its assumptions, prejudices and biases, and to encourage the development of arguments to address them. It is also important to remember that parents, teachers and school communities tend to be seen as key audiences for their role in exposing and protecting children from content that incites hate, as well as those who have the ability to shape the legal and political landscape of online hate speech, including policymakers and NGOs, and those who can have a significant impact on online communities by denouncing hate speech, particularly journalists, bloggers and activists (Gagliardone et al. 2015).

Conclusions

Education is one of the most important ways in which we can address and prevent the harms of hate speech (Molnar, 2012). Therefore, both the methodology and content of a counter-intervention were chosen, taking into account the findings of both an experimental study and a systematic literature review.

Firstly, regarding the content a common denominator of the analyzed initiatives is the emphasis on the development of critical thinking skills and the ethically-reflective use of social media as starting points of media and information literacy skills to combat hate speech online. The expectation is that these media and information literacy competencies can enhance individuals' ability to identify and question hateful content online, understand some of its assumptions, biases and prejudices, and encourage the elaboration of arguments to confront it.

Global citizenship education is an effective model in this sense, as it is based on building an awareness of the dignity of all human beings, a sense of belonging to a global community and people's involvement, both as individuals and collectively, in order to drive cultural, social and political change for the construction of a more just and sustainable world. Many experts believe the core concepts of global citizenship education (GCED) can play an important role in countering hate speech (Gagliardone et al., 2015). Through the promotion of global skills education, individuals can develop the knowledge and skills needed to combat hate speech, empowering students to fulfil their responsibilities and create a more just and inclusive society (Andreotti, 2006). In fact, this educational approach allows for critical thinking on complex global matters, motivating individuals to learn, voice their thoughts, make informed decisions, and actively contribute to building a fairer and more sustainable world. Nevertheless, the target audience of each initiative may determine specific content in order to achieve three common educational objectives: to educate, to examine, and to counteract online hate speech.

Secondly, in view of these training contents and the complex intervention methodology outlined, the characterization of the intervention within the framework of an educational community program is considered opportune. The latter, in its renewed conception, is closely linked to the phenomenon of "educational poverty", which urges a 360-degree view of the growth processes of children and adolescents, and has proven effective, for example, in promoting gender equality through the development of knowledge, skills, values and attitudes that promote equality between women and men, develop respect and enable young people to question gender-based expectations and roles, which is, for example, one of the fundamental themes of global citizenship education (Sant et al., 2018). With this premises, the community education is evaluated as the most effective methodology considering the assumption of a theoretical background in which learning occurs through engagement in authentic experiences involving active manipulation and experimentation with ideas and artefacts, rather than through the accumulation of static knowledge (Bruner, 1973). In fact, it provides a responsive, community-based system for collective action by all educational and community agencies to address community issues.

Lastly, the structure of the intervention should replicate and adapt the one already developed by Wachs and colleagues (2023) in the context of HateLess project, with five modules (one per week) of 90 minutes delivered in blended mode, to meet the different training needs of the members of the learning community.

References

Andreotti, V. (2006) Soft versus critical Global Citizenship Education, Policy & Practice – A Development Education Review, issue 3, 40–51. http://www.developmenteducationreview.com/issue3-focus4

Bagnato, K. (2020), Online hate speech: responsabilità pedagogico-educative. Analisi online della Didattica e della Formazione Docente, 12(20), 195-211;

Bilewicz M., Soral W., Marchlewska M., Winiewski M. (2015). When Authoritarians Confront Prejudice. Differential Effects of SDO and RWA on Support for Hate-Speech Prohibition; University of Warsaw;

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;

Boag, E. M., & Carnelley, K. B. (2016). Attachment and prejudice: The mediating role of empathy. British Journal of Social Psychology, 55(2), 337–356. https://doi.org/10.1111/bjso.12132;

Bortone R., Cerquozzi F. (2017). L'hate speech al tempo di Internet; Università degli Studi Roma Tre. FCSF - Aggiornamenti Sociali;

Bruner, J. S, (1973). *Beyond the information given*. New York: Norton

Celuch, M., Oksanen, A., Räsänen, P., Costello, M., Blaya, C., Zych, I., Llorent, V. J., Reichelmann, A., & Hawdon, J. (2022). Factors Associated with Online Hate Acceptance: A Cross-National Six-Country Study among Young Adults. International Journal of Environmental Research and Public Health, 19(1), 534 https://doi.org/10.3390/ijerph19010534;

De Caroli M. E. (2016) *Categorizzazione sociale e costruzione del pregiudizio. Riflessioni e ricerche sulla formazione degli atteggiamenti di «genere» ed «etnia»*, Milano, Franco Angeli, 2016

European Commission (2020), Communication from the commission to the european parliament and the council, a more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime, Brussels;

Floridi, 2014, *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*, Oxford: Oxford University Press

Gagliardone I., Gal D., Alves T., Martinez G. (2015), *Countering online hate speech*, Unesco, Paris;

Hornsby J. (2003) *Free Speech and Hate Speech: Language and Rights*, Quodlibet, Normatività Fatti Valori

La Benevolencija (2005 a), Stichting Radio La Benevolencija/Humanitarian Tools Foundation (Radio Benevolencija). http://www.labenevolencija.org.;

La Benevolencija (2005 b), Great Lakes Reconciliation Radio. Radio programs and grassroots activities as instruments of prevention of conflict, reconciliation and restoration in the Great Lakes region (Rwanda, Burundi, Kivu - DRC), Amsterdam;

Marinelli A. (2021). *Educare alla cittadinanza digitale nell'era della platform society*, il Mulino, Bologna.

Molnar, P. (2012) Responding to "Hate Speech" with Art, Education, and the Imminent Danger Test. in The Context and Content of Hate Speech: Rethinking Regulation and Response, ed. Michael Herz and Peter Molnar (New York: Cambridge University Press, 2012), 183—197.

Paluck E. L. 2006. The second year of a "New Dawn." Year two evidence for the impact of the Rwandan reconciliation drama Musekeweya. Unpublished evaluation report for La Benevolencija;

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. European Journal of Social Psychology, 38(6), 922–934. https://doi.org/10.1002/ejsp.504;

Prezza, M., Trombaccia, F.R., & Armento, L. (1997). La scala dell´autostima di Rosenberg: traduzione e validazione italiana. *Bollettino di Psicologia Applicata*, 223, 35–44

Rivoltella, P. C., Rossi, P. G., & Aroldi, P. (2019). *Tecnologie per l'educazione*. Pearson.

Senato della Repubblica (2022) Analisi comparativa sul fenomeno dell'istigazione all'odio online;

Straub E.., Pearlman L.A., Gubin A., Hagengimana A. (2005), Healing, reconciliation, forgiveness and the prevention of violence after genocide or mass killing: An intervention and its experimental evaluation in Rwanda, Journal of Social and Clinical Psychology 24: 297-34;

Van Hiel, A., & De Clercq, B. (2009). Authoritarianism is good for you: Right-wing authoritarianism as a buffering factor for mental distress. European Journal of personality, 23, 33–50. doi:10.1002/per.702;

Vollhardt J., Coutin M., Staub E., Weiss G. and Deflander J (2006). *Deconstructing Hate Speech in the DRC: A Psychological Media Sensitization Campaign*, University of Massachusetts, Amherst;

Wachs S, Machimbarrena JM, Wright MF, Gámez-Guadix M, Yang S, Sittichai R, Singh R, Biswal R, Flora K, Daskalou V, Maziridou E, Hong JS, Krause N. (2022). Associations between Coping Strategies and Cyberhate Involvement: Evidence from Adolescents across Three World Regions. Int J Environ Res Public Health. doi: 10.3390/ijerph19116749;

Wachs, S., Castellanos, M., Wettstein, A., Bilz, L., & Gámez-Guadix, M. (2023). Associations Between Classroom Climate, Empathy, Self-Efficacy, and Countering Hate Speech Among Adolescents: A Multilevel Mediation Analysis. *Journal of Interpersonal Violence*, 38(5–6), 5067–5091. https://doi.org/10.1177/08862605221120905

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