Online Journal of Communication and Media Technologies, 2023, **13**(4), e202344
e-ISSN: 1986-3497

OPEN ACCESS

**Research Article**

# Effect of online civic intervention and online disinhibition on online hate speech among digital media users

**Shuaa Aljasir** [1*]
0000-0002-1165-7900

[1] King Abdulaziz University, Jeddah, SAUDI ARABIA
[*] Corresponding author: shaljasir@gmail.com

**Citation:** Aljasir, S. (2023). Effect of online civic intervention and online disinhibition on online hate speech among digital media users. *Online Journal of Communication and Media Technologies, 13*(4), e202344. https://doi.org/10.30935/ojcmt/13478

**ARTICLE INFO**

Received: 23 Mar 2023
Accepted: 4 Jul 2023

**ABSTRACT**

Given the numerous theoretical gaps in explanations regarding online hate speech and the insufficient level of empirical data to fully understand this phenomenon, this study utilized an explanatory sequential mixed-method approach through two phases; it employed a quantitative online questionnaire (n=2,242), followed by a qualitative online vignette interview (n=23) to contribute to the knowledge in the field. In particular, it aimed to investigate the moderating roles of online civic intervention (OCI), online disinhibition, and demographic variables in the relationship between online hate exposure\victimization and perpetration. Among the most interesting findings of this research was that the impact of toxic online disinhibition was a negative moderator in the relationship between online hate exposure and perpetration. Furthermore, the impact of high-threshold OCI was positive in this relationship. However, the impact of low-threshold OCI was positive in the relationship between online hate speech victimization and perpetration. Further justifications for OCI and new proposed types of online disinhibition have been suggested based on the findings.

**Keywords:** online civic intervention, online disinhibition, online hate, hate speech, digital media

## INTRODUCTION

The Internet has enabled global forms of interaction in which people may express themselves freely and have their messages magnified, including those who promote hatred. The incidence of online hate speech has increased exponentially and is now accessible to a worldwide audience free of charge and without mediation (Beausoleil, 2019). In recent years, the expansion and development of digital media tools appear to have contributed to its pervasiveness. Many digital media users around the world have admitted to facing hate speech in recent years (Schmid et al., 2022). Hate speech, especially in today's digital society, can have direct and\or indirect negative consequences for those targeted (Chetty & Alathur, 2018). Moreover, it can also have adverse psychological, social, and legal effects. Prior studies have shown that exposure to online hate speech can contribute to numerous psychological and psychiatric disorders, including depression, anxiety, stress, sleep disorders, and a feeling of insecurity (Keipi et al., 2016; Näsi et al., 2015; Saha et al., 2019; Wachs et al., 2022). Focusing on the negative social impacts of exposure to online hate speech, some studies have found such impacts to be closely associated with a weak family network/bond, a poor-quality social life and experience of traditional bullying (Oksanen et al., 2014). As the law concerning hate speech is not always sufficiently clear that average individuals can understand it, individuals in countries worldwide may find it difficult to determine if their actions fall outside what is permitted by the law, which may place them at risk of adverse legal consequences due to inadvertently creating hateful content (Brown, 2018). While it has previously been demonstrated that being subject to online hate speech can be devastating and have long-lasting consequences (Saha et al., 2019), recent research has revealed there to be a number of theoretical gaps concerning the clarification of related conduct as well as insufficient empirical data to fully comprehend such conduct (Castaño-Pulgarín et al., 2021). Thus, to reduce the occurrence of online hate speech, it is vital to identify the online features and factors that encourage or prevent its dissemination. However, few predictors of online hate speech have so far been identified (Celuch et al., 2022; Wachs et al., 2022).

In addition, it must be recognized that hate speech can have multiple manifestations that can be viewed, acknowledged, and/or understood in different ways by individuals from diverse backgrounds and who exhibit differing normative conceptions (Bormann et al., 2021). Yet, most prior studies concerning hate speech have been conducted in English-speaking cultural contexts (Rizwan et al., 2020), meaning that there has been insufficient research on this highly relevant contemporary phenomenon and its consequences in other cultures. In terms of the Arab cultural context, research has confirmed that individuals' behaviors, actions, and communications are all strongly influenced by the culture and its norms (Obeidat et al., 2012). For instance, Hofstede et al. (2011) described Arab nations as having a wide power distance, a moderate divide between masculinity and femininity, somewhat strong uncertainty avoidance, and high collectivism. Moreover, because the majority of individuals living in Arab nations are Muslims, their cultural values are founded on Islamic principles, which is evident in their speech, actions, social interactions, and behaviors during other activities (Anggraeni & Tajuddin, 2018). In contrast to the practice in English-speaking societies, in Arab cultures speech appears to be a social instrument employed in the construction of society. Indeed, according to Zaharna (1995), emphasis is placed on form over function, effect over accuracy, and image over message. Hence, individuals in Arab societies appear to employ speech that is more direct when it comes to conveying their intended expression or feeling. In fact, research suggests that the direct method of expression is the method most frequently employed in Arab societies (Anggraeni & Tajuddin, 2018). These unique characteristics of the Arab cultural context render it fundamental to investigate the ways in which individuals belonging to this culture are exposed to, are victimized by, and perpetrate online hate speech via digital media platforms.

To bridge this research gap, the present study is one of the first to examine the preventive and promotive roles played by online civic intervention (OCI), online disinhibition, and various demographic variables as moderators of the relationship between online hate exposure and perpetration among Arab users of digital media.

# THEORETICAL BACKGROUND

## Online Hate Speech

### *Defining online hate speech*

According to Leets and Giles (1999), hate speech is a subcategory of violent conversation, defined as statements designed to impose harm and, independent of purpose, which are perceived to do harm. Hate speech is the distribution, provocation, or advocacy of hatred, violence, and discrimination against a person or group based on their traits, such as race, religion, color, gender, nationality, political affiliation, and disability (Zhang & Luo, 2018). Thus, online hate speech may be described as vocal demonstrations of hatred in online environments, generally employing derogatory, degrading, and threatening language (Meibauer, 2013).

### *Online vs. face-to-face hate speech*

UNESCO cited four significant distinctions between online and face-to-face hate speech. The first is the persistence of online hatred, that is, its potential to stay "active" over extended periods and in several mediums. The second distinction is the unexpected recurrence of online hate speech, which might reemerge elsewhere even if it has been deleted from one location. The third distinction is the significance that anonymity acquires online: It is a privilege that permits, under certain situations, the disclosure of information that one would otherwise be reluctant to reveal. The fourth distinction is transnationality, which enhances the impact of hate speech and makes identifying legal measures to address it more difficult (Gagliardone et al., 2015).

### *Online hate speech and demographic variables*

Due to the numerous distinguishing qualities and characteristics of their users, digital media platforms are considered the ideal location for the dissemination of online hate speech (Schmid et al., 2022). A review of the literature revealed that prior studies have investigated differences in online hate involvement based on the roles of individual demographic variables and reported contradictory findings. For example, focusing on users' age, a study conducted in four European nations by Hawdon et al. (2017) found that an increasing age increases the probability of exposure to online hate among British and American users of digital media platforms. By contrast, Kaakinen et al. (2018) focused on samples from the same nations and found no significant relationship between age as a control variable and the production of hate. The literature concerning the role of gender in the production of or exposure to online hate has shown that men tend to produce more online hate content, whereas women tend to avoid creating such materials. For instance, Costello and Hawdon (2018) reported that males produce more online hate content than females. In line with this finding, a study by Wilhelm and Joeckel (2019) revealed that females are more considerate when it comes to avoiding producing harmful content than males. However, Hall (2018) found that age and gender did not significantly predict exposure to online hate content. On the contrary, with regard to education level, Hall's (2018) study revealed that a higher education level significantly predicts exposure to online hate content, which accords with the findings of Milani et al. (2022). By contrast, Näsi et al. (2021) showed that an individual's education level does not have a significant effect on their risk of exposure to online harassment. Despite such findings, no previous study has tested whether the interactions among these demographic variables influence their moderating role in the relationship between exposure to and perpetration of online hate speech. Thus, this study aims to answer the following research question and hypothesis:

**RQ1.** To what extent are Arab users exposed to, victimized by, and perpetrate online hate speech through digital media platforms?

**H1.** Users' demographic variables (age, gender and educational level) moderate the relationship between online hate speech exposure\victimization and perpetration.

**H1a.** Users' demographic variables (age, gender and educational level) moderate the relationship between online hate speech exposure and perpetration.

**H1b.** Users' demographic variables (age, gender and educational level) moderate the relationship between online hate speech victimization and perpetration.

## Online Disinhibition

### *Defining online disinhibition*

Suler (2004) was the first to propose online disinhibition as a phenomenon in which users in cyberspace react in a different way from their face-to-face interaction as they feel less compelled and are capable of expressing themselves more freely.

### *Dimensions and aspects of online disinhibition*

The literature documents two dimensions of online disinhibition: benign disinhibition, which occurs when people are driven to reveal personal information, concealed emotions, concerns, and desires (Lapidot-Lefler & Barak, 2015), and toxic disinhibition, which is comparable to the "troll" of today (Wachs et al., 2022).

Suler highlighted six frequently examined aspects of online disinhibition:

(1) dissociative anonymity, which is defined as "the degree to which a person thinks that they may conceal or modify their genuine identity in the online world",

(2) "invisibility, which is "degree to which a person believes that others in online world cannot see them",

(3) asynchronicity, which is defined as "the degree to which a person feels that the form of communication facilitates delayed replies in the online world",

(4) solipsistic introjection, which is "the extent to which an individual in online conversation senses a voice or picture of the other person in their thoughts",

(5) dissociative imagination, which is "the extent to which a person views the online environment as an unreal, fictitious world", and

(6) minimization of authority, which is believed to be "the degree to which a person sees the absence or declining influence of real-life authority figures in the online world" (Cheung et al., 2016).

### *Online disinhibition and online hate speech*

Reviewing the literature showed that very few studies have investigated the possible moderating role played by online disinhibition. For instance, a study conducted by Wachs and Wright (2018) only investigated the moderating role of toxic disinhibition in the association between bystanders of online hate and perpetration. It revealed that bystanders indicated higher online hate crimes and greater online disinhibition levels. Another study conducted by some of the same researchers investigated its role between cyberbullying and cyberhate perpetration and revealed similar results (Wachs et al., 2019). Conversely, Harriman et al. (2020) confirmed the association between exposure to hate messages and benign disinhibition. The current research contributes to the existing knowledge by investigating the moderating role of each of the online disinhibition dimensions in the relationship between online hate speech exposure\victimization and perpetration. It does so by testing the following hypothesis and sub-hypotheses:

**H2.** Online disinhibition moderates the relationship between online hate speech exposure\victimization and perpetration.

**H2a.** Benign online disinhibition moderates the relationship between online hate speech exposure and perpetration.

**H2b.** Benign online disinhibition moderates the relationship between online hate speech victimization and perpetration.

**H2c.** Toxic online disinhibition moderates the relationship between online hate speech exposure and perpetration.

**H2d.** Toxic online disinhibition moderates the relationship between online hate speech victimization and perpetration.

## Online Civic Intervention

### *Defining online civic intervention*

Whereas in conventional media environments consumers are reliant on organizations to act against harmful information, digital media platforms enable ordinary individuals to counter the problematic material they detect (Lim & Golan, 2011). Interference tactics in the digital world generally suggest that "the state as a regulatory agency legislates"; "platform operators" and users can interfere against material viewed as undesirable, or what is known as "online civic intervention (OCI)" (Riedl et al., 2020). OCI is defined as activity performed by regular users to combat troublesome online conduct with the goal of reestablishing civil and logical public dialogue (Kunst et al., 2021).

### *Types of online civic intervention*

According to the literatures, users prefer one of two types of OCI: low-threshold OCI, which employs media platform mechanisms, such as reporting offensive speech to online platform owners and grading individuals' statements using platform-based buttons. Conversely, high-threshold OCI involves direct participation in civil counter speech and encouraging people who generate hate speech or derogatory language to be courteous of others (Porten-Cheé et al., 2020).

### *Online civic intervention and online hate speech*

Reviewing the literature revealed very few studies focusing on OCI and its relation to online hate speech. For instance, Kunst et al. (2021) investigated whether solidarity citizenship norms promote OCI in general. The outcomes demonstrated that users who maintain unity citizenship standards are more inclined to flag hate commentary and be involve in counter speech. Obermaier (2022) found that perceived individual responsibility for preventing online hate speech predicted greater direct and indirect involvement from online bystanders. Additionally, regular exposure to online hate speech was affiliated with prevention and intervention. As little is known about why individuals prefer one kind of OCI over the other and the role it could play as a moderator in minimizing online hate perpetration, this research tests the following hypothesis:

**H3.** OCI moderates the relationship between online hate speech exposure\victimization and perpetration.

**H3a.** Low-threshold OCI moderates the relationship between online hate speech exposure and perpetration.

**H3b.** Low-threshold OCI moderates the relationship between online hate speech victimization and perpetration.

**H3c.** High-threshold OCI moderates the relationship between online hate speech exposure and perpetration.

**H3d.** High-threshold OCI moderates the relationship between online hate speech victimization and perpetration.

## METHOD

### Research Design and Participants

To fully accomplish its aims, an explanatory sequential mixed-method approach through two phases was utilized. First phase employed a quantitative online questionnaire across several popular accounts on digital media platforms to recruit participants. Participants of the phase consisted of 2,242 voluntary individuals aged between 18 and 65 years (mean [M]=30.46, standard deviation [SD]=9.78), of whom 1,583 were male (70.6%) and 659 were female (29.4%). Level of education was measured on a scale from 1 to 5 (1=low level of education), with an average education level of high school (SD=1.20). After filling in questionnaire, participants were asked whether they wanted to participate in the second qualitative phase, an online vignette interview, by writing down their contact information. There were 23 second-phase participants of which nine were male and 14 were female. Their ages ranged between 18 to 42 years (M=28.80, SD=7.60).

### Measures

After enquiring about the participants' demographic information (i.e., age, gender, and educational level), they were asked about their online hate involvement (exposure, victimization, and perpetration) following the method of Hawdon et al. (2017): How often in the past 12 months have you:

(1) "personally been the target of hateful or degrading content on digital media platforms because of your sex, religious affiliation, or race?",

(2) "been exposed to hateful or degrading content related to sex, religious affiliation, or race on digital media platforms?", and

(3) "posted hateful or degrading content on digital media platforms, which inappropriately attacked certain groups of people or individuals based on their sex, religious affiliation, or race?"

The participants rated each item on a scale from 0 (never) to 4 (very frequently). To measure online disinhibition, a scale proposed by Udris (2014) was adopted. It comprised 11 items entailing two subscales: benign disinhibition (seven items, e.g., "It is easier to connect with others online than talking in person") and toxic disinhibition (four items, e.g., "I do not mind writing insulting things about others online, because it's anonymous"). The items were measured on a 7-point Likert scale, with the participants indicating how likely they agreed (1=totally disagree; 7=totally agree). To measure the two dimensions of OCI, a scale by Kunst et al. (2021) was adopted, with two items measuring low-threshold OCI (e.g., "I would report user comments to the platform if there was an option to flag") and four items measuring high-threshold OCI (e.g., I would call upon other users to report the comment to the platform operator."). The items were measured on a 7-point Likert scale, with the sample stating how likely they were to involve in the mentioned behavior if they encountered online hate content written by other users (1=very unlikely; 7=very likely). To ensure the content validity of the scales among digital Arab users, the scales were evaluated by two independent experts who hold PhD degrees to assess their comprehensibility and suitability for the target group.

### Interviews

Using the Tweetgen generator, the participants were presented with three fictional online hate speech posts on race, gender, and disability on Twitter–which is among the most used digital media platforms among Arabs (Global Media Insight, 2022). These posts were chosen because of a suggestion by Wilhelm et al. (2020) that offensive language targeting social groupings is seen as highly dangerous. Five trial interviews were

**Table 1.** Descriptive & reliability measures of scales of study

| | Subscale | n | A | Mean | Standard deviation | Kurtosis | Standard error |
|---|---|---|---|---|---|---|---|
| Online disinhibition | Benign | 2,242 | .957 | 4.0108 | 1.74662 | -1.551 | .103 |
| | Toxic | 2,242 | 2,242 | 2.4312 | 1.33060 | 1.131 | .103 |
| Online civic intervention | Low threshold | 2,242 | 2,242 | 3.8100 | 1.97751 | -1.670 | .103 |
| | High threshold | 2,242 | 2,242 | 2.4301 | 1.26105 | 1.267 | .103 |
| Hate victimization | - | 2,242 | 2,242 | 2.2337 | 0.01697 | 0.089 | .103 |
| Hate exposure | - | 2,242 | 2,242 | 2.1039 | 0.86963 | -0.332 | .103 |
| Hate perpetration | - | 2,242 | 2,242 | 1.8702 | 1.05604 | -0.727 | .103 |

conducted to verify that the vignettes depicted common and realistic scenarios. Based on the input, they were modified further. After reading the posts, the participants were asked to evaluate whether they considered what was written as hateful speech. They were asked to recall and remark on comparable circumstances that they had experienced or witnessed as digital media platform users. To further investigate the role of online disinhibition, the participants were asked to state the reasons behind posting such hateful content online. Finally, to clarify more on the role of OCI, they were asked what they would do if they were to encounter such a situation. The interviews, conducted via the Zoom video-conferencing platform, lasted approximately 25 minutes and were transcribed verbatim, then analyzed using thematic analysis.

## Data Analysis

### *Study one*

The dataset under consideration consisted of descriptive data collected from 2,242 participants (**Table 1**). To determine the reliability and internal consistency of the survey instrument, Cronbach's alpha coefficient was calculated. The acceptability threshold for Cronbach's alpha typically ranged from 0.7 to 0.9, indicating a desirable level of internal consistency. In this dataset, the calculated Cronbach's alpha coefficient was between 0.8 and .957, surpassing the lower limit of acceptability. This indicated that the questionnaire items were reliably measuring the underlying construct of job satisfaction. Furthermore, the dataset's kurtosis values were examined to assess the normality of the distribution of variables. With a general guidance suggesting that values within the range of -2 to +2 were considered acceptable, indicating a roughly normal distribution. It was found that the kurtosis values for the variables of interest ranged from -1.670 to +1.131. These values fell within the acceptable range, suggesting that the distributions of the variables were approximately normal. This indicated that the regression analysis could assume normality and rely on assumptions associated with such distributions.

### *Study two*

This study applied the qualitative thematic data analysis process proposed by Lester et al. (2020) to analyze the data gathered through the interviews. More specifically, after transcribing the interviews verbatim, an initial analysis was conducted to allow the researcher to become familiar with the data, while memos were produced to provide a primary explanation of the data and elucidate any developing explanations. The data were then coded, and the codes were aggregated into categories that were used to produce themes. The generated themes accorded with the conceptual and analytical goals of the present study. To ensure the intercoder reliability and the agreement of the data analysis, this study followed the recommendation of Campbell et al. (2013) and asked two independent researchers to code three interview transcripts (representing more than 10% of the total number of transcripts).

# RESULTS

## Study One

### *H1. Users' demographic variables (age, gender, and educational level) moderate the relationship between online hate speech exposure\victimization and perpetration*

To explore the impact of the participants' age, gender, and education as moderators of the relationship between hate speech exposure and perpetration, an initial investigation was carried out whereby interaction

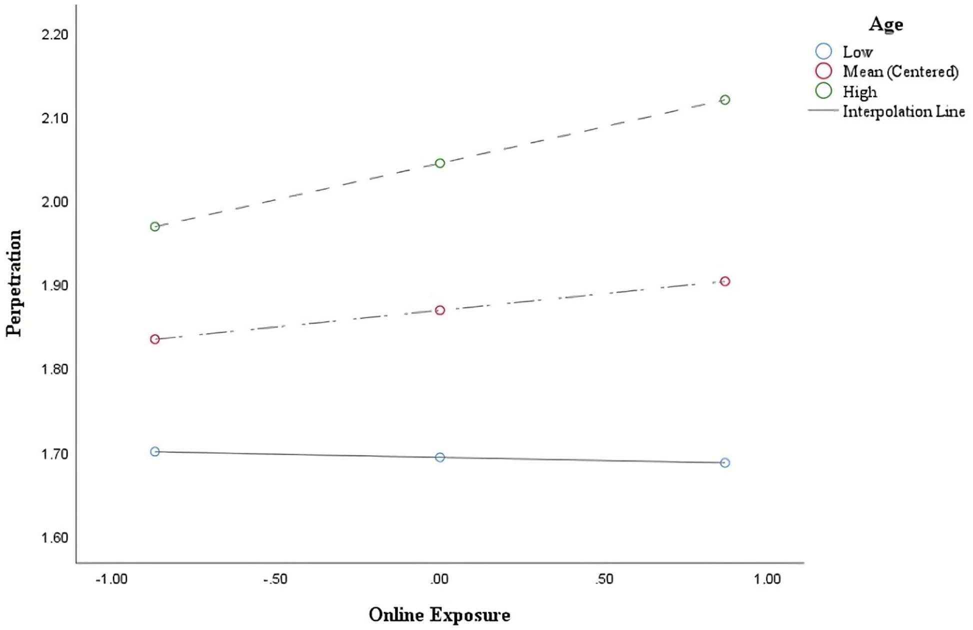

**Figure 1.** Plot of conditional effects of age as a moderator of relationship between online exposure & perpetration at low, medium, & high values (Source: Author)

terms relating to hate speech exposure were created for all three demographic variables. Regression analysis was then performed in search of potential moderators. From this preliminary analysis, education and gender were rejected as non-significant interactions (p≥.05). Moderation analysis was then carried out using Hayes process v.4.1. for SPSS, model one, with age as the potential moderator between hate speech exposure and hate speech perpetration. Overall, the model was significant (F[3,2238]=24.1232, p<.001; $R^2$=.1770). The results of the analysis demonstrate that there was a positive and significant moderating impact of age on the relationship between hate speech exposure and hate speech perpetration (b=.0048, CI[.0002, .0094], p<.05; $R^2$ change=.0018). Furthermore, a simple slope analysis (**Figure 1**) demonstrated that the moderation effect of age grew stronger as education increased. At lower ages, the conditional effect=-.0076 (CI[-.0789, .0637], p=.8342). At a medium age, the conditional effect=.0397 (CI[-.0103, .0897], p=.1194). At higher ages, the conditional effect=.0870 (CI[.0242, .1499], p<.001). Therefore, changes in age increased the strength of the relationship between online exposure and perpetration. The simple slope analysis allowed the additional analysis to explore in what age group the effect was strongest, age was split into relative age groupings. This analysis demonstrated that for relatively older participants, as opposed to the younger age groups, there was a significant relationship between exposure and perpetration.

Age, gender, and education were also explored as potential moderators of the relationship between hate speech victimization and perpetration. A similar procedure was employed whereby the interaction with the demographic variables was initially explored using regression analysis prior to performing moderation analysis using Hayes Process. This process uncovered age and gender as non-significant interactions (p≥.05) and were therefore omitted from the model. Overall, the model was significant (F[3,2238]=57.5544, p<.001; $R^2$=.2676). The results of the moderation analysis showed a positive and significant moderating effect of education on the relationship between hate speech victimization and hate speech perpetration (b=.0546, CI[.0192, .0899], p<.01; $R^2$ change=.0038). A simple slope analysis (**Figure 2**) demonstrated a significant effect of education on all values, with an increasing moderating effect as the variable increased. Therefore, at low levels of education, the conditional effect=.2718 (CI[.2077, .3360], p<.001). At a medium level of education, the conditional effect=.3375 (CI[.2847, .3903], p<.001). At higher levels of education, the conditional effect=.4032 (CI[.3318, .4745], p<.001). These results suggest that generally, changes in education scores affected the relationship between online exposure and perpetration. When examined over relative levels of high medium and low scores, this effect was present in all groups, but it was observed that as education levels increased, the effect of this moderation increased as well.

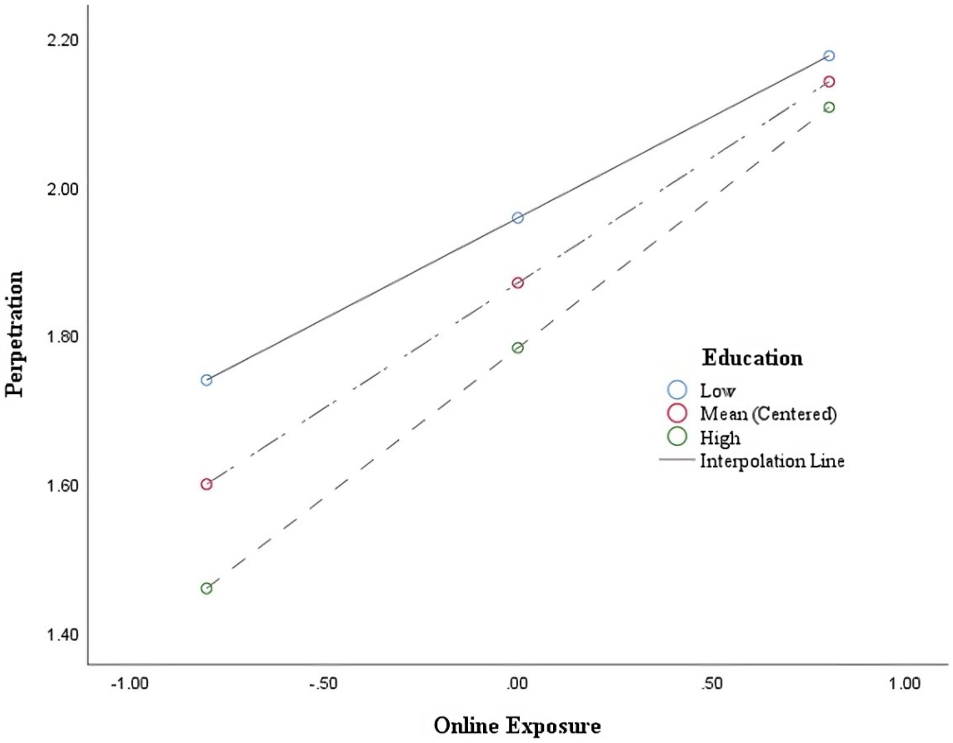

**Figure 2.** Plot of conditional effects of education as a moderator of relationship between online victimization & perpetration at low, medium, & high values (Source: Author)

### *H2. Online disinhibition moderates the relationship between online hate speech exposure\victimization and perpetration*

The effects of both benign and toxic online disinhibition as moderators of online hate speech exposure and perpetration (**H2a** & **H2c**) were explored in the same model. Both the benign and toxic scores were entered into model two of Hayes process v.4.1. for SPSS. The independent variable was online hate speech exposure, and the dependent variable was perpetration. Overall, the model was significant (F[5,2236]=259.1463, p<.001; $R^2$=.6057). The impact of benign online disinhibition as a moderator was not significant (b=-.0253, CI[-.0541, .0035], p=.0848; $R^2$ change=.0008). Conversely, toxic online disinhibition was a significant negative moderator of the relationship between online hate speech and perpetration (b=-.0253, CI[-.0541, .0035], p=.0848; $R^2$ change=.0008). Taken together, the $R^2$ change was significant (p<.001). Therefore, hypothesis **H2a** was rejected, and hypothesis **H2c** was supported. A simple slope analysis (**Figure 3**) demonstrated that at low values, the toxic inhibition scores were not a significant moderator of low, mean, and high benign inhibition scores (p>.05). However, its effect as a moderator became significant at high levels of toxic disinhibition (p<.01). This analysis demonstrated that changes in benign disinhibition scores did not reflect changes in the strength of the relationship between online exposure and perpetration. However, in the case of toxic disinhibition changes in that score changed the strength of the relationship between online exposure and perpetration.

To examine the effects of benign and toxic disinhibition as moderators on the relationship between online victimization and perpetration, model two of Hayes process v.4.1. for SPSS was used. The independent variable was online hate speech victimization, and the dependent variable was perpetration. Benign and toxic disinhibition were entered as two separate moderators. Overall, the model was significant (F[5,2236]=262.6741, p<.001; $R^2$=.3700). However, the impact of benign online disinhibition as a moderator was not significant (b=-.0437, CI[-.0032, .0907], p=.0681; $R^2$ change=.0009). Further, toxic online disinhibition was a non-significant moderator of the relationship between online hate victimization and perpetration (b=-.0010, CI[-.0457, .0436], p=.9636; $R^2$ change=.0000). Therefore, hypotheses **H2b** and **H2d** were both rejected. In this analysis, change in scores of neither benign nor toxic disinhibition resulted in changes to the relationship between online hate victimization and perpetration.

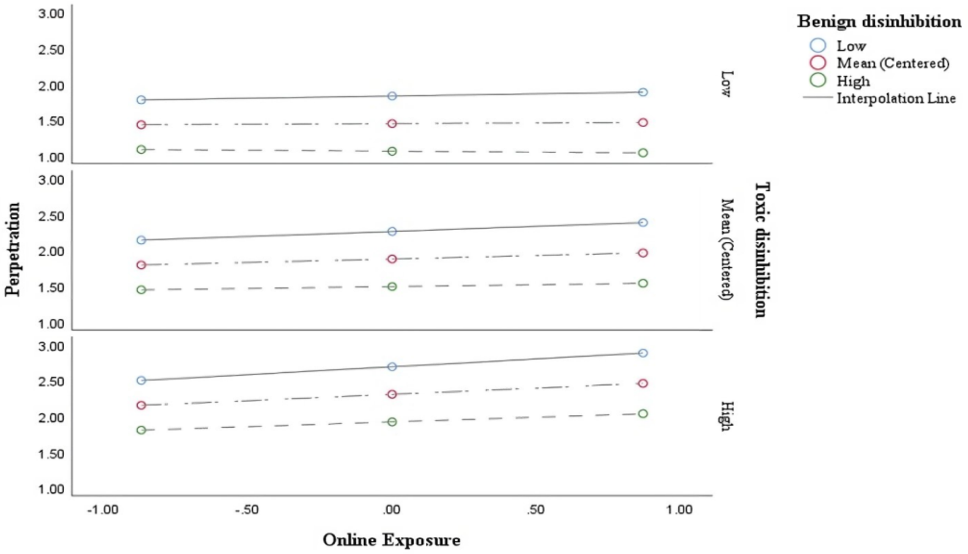

**Figure 3.** Plot of conditional effects of benign disinhibition as a moderator of relationship between online exposure & perpetration paneled at low, medium, & high values of toxic disinhibition (Source: Author)

### *H3. OCI moderates the relationship between online hate speech exposure\victimization and perpetration*

The moderating effects of low- and high-threshold OCI in the relationship between online hate speech exposure and perpetration (**H3a** & **H3c**) were examined using model two of Hayes process v.4.1. for SPSS. The independent variable was online hate speech exposure, and the dependent variable was perpetration, with low- and high-threshold OCI entered separately as potential moderators. Overall, the model was significant (F[5,2236]=236.4712, p<.001; $R^2$=.5881). The impact of low-threshold OCI as a moderator was not significant (b=-.0012, CI[-.0248, .0271], p=.9308; $R^2$ change=.0000). However, the impact of high-threshold OCI as a moderator was significant and positive (b=.0346, CI[.0017, .0675], p<.05; $R^2$ change=.0012). Taken together, the $R^2$ change was significant ($R^2$=.0014, p<.001). Thus, hypothesis **H3a** was rejected, and hypothesis **H3c** was supported. A simple slope analysis (**Figure 4**) showed that high-threshold OCI was a significant moderator at low, mean, and high values (p<.05), except in single circumstance of high scores in low-threshold OCI combined with high values in high-threshold OCI (p=.1504). These results suggest that generally, changes in high threshold OCI scores affected relationship between online exposure and perpetration. When examined over relative levels of high medium and low scores, this effect was present in all groups.

The moderating effect of low- and high-threshold OCI in the relationship between online hate speech victimization and perpetration (**H3b** & **H3d**) was also examined using model two of Hayes Process v.4.1. for SPSS. The independent variable was online hate speech victimization, and the dependent variable was perpetration, with low- and high-threshold OCI entered separately as potential moderators. Overall, the model was significant (F[5,2236]=253.2400, p<.001; $R^2$=.6013). Impact of low-threshold OCI as a moderator was positive and significant (b=1190, CI[.0803, .1576], p<.01; $R^2$ change=.0104). However, the impact of high-threshold OCI as a moderator was not significant (b=-.0213, CI[-.0622, .0196], p=.3073; $R^2$ change=.0003). Taken together, the $R^2$ change was significant ($R^2$=.0126, p<.01). Therefore, hypothesis **H3b** was supported, and hypothesis **H3d** was rejected. A simple slope analysis (**Figure 5**) demonstrated that low-threshold OCI was a significant moderator at low, mean, and high scores (p<.05), except in the single circumstance of low scores in high OCI and high values in high-threshold OCI (p=.1504). These results suggest that generally, changes in low threshold OCI scores affected the relationship between victimization and perpetration. When examined over relative levels of high medium and low scores, this effect was present in all groups.

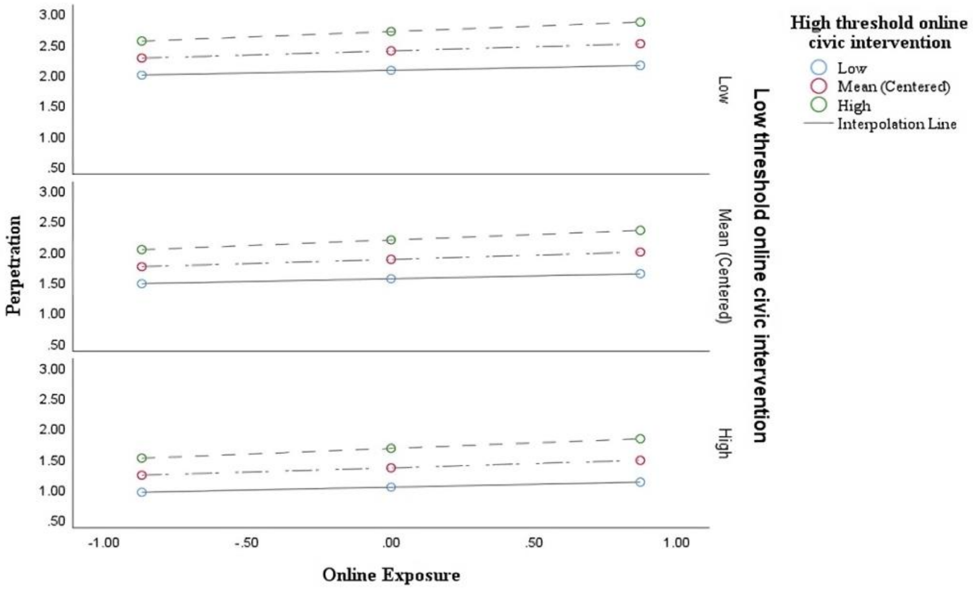

**Figure 4.** Plot of conditional effects of high-threshold OCI as a moderator of relationship between online exposure & perpetration paneled at low, medium, & high values of low-threshold OCI (Source: Author)

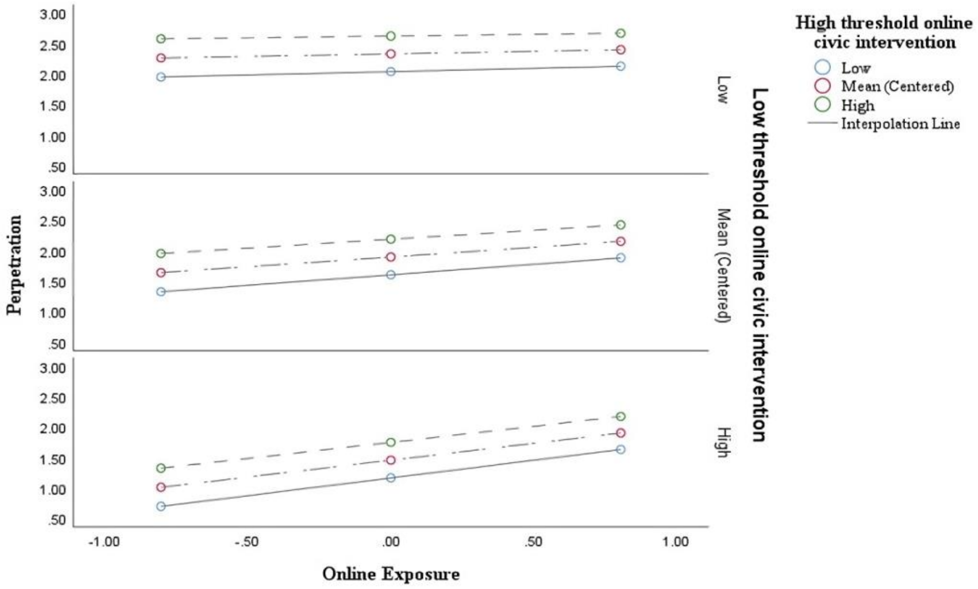

**Figure 5.** Plot of conditional effects of high-threshold OCI as a moderator of relationship between online victimization & perpetration paneled at low, medium, & high values of low-threshold OCI (Source: Author)

## Study Two

The interviews with the participants about their experience regarding online hate speech began with a presentation of three fictional online hate speech posts on Twitter. Most of the participants considered the posts to be hateful (**Table 2**).

However, a number of them (mostly male) thought that the posts were not hateful and could be considered either as expressing someone's true feelings–as the Internet allows greater freedom to say whatever they thought compared to real-world communication–or a reaction to very rude mistreatment that a user might have faced, leading them to behave in this way.

The participants were asked to recall experiences similar to those in the tweets. Almost all of them indicated that they had been exposed to such content, and few of them indicated that they had been victims of this kind of speech. The type of hateful speech they mentioned as occurring most frequently were in

**Table 2.** Participants' reactions to fictional online hate speech posts on Twitter used

| | Hateful | n | Males | Females |
|---|---|---|---|---|
| Race | Yes | 21 | 7 | 14 |
| | No | 2 | 2 | 0 |
| Gender | Yes | 21 | 8 | 13 |
| | No | 2 | 1 | 1 |
| Disability | Yes | 20 | 7 | 13 |
| | No | 3 | 2 | 1 |

relation to religion, nationality, and gender. For instance, disrespecting users' beliefs and gender were the most recurring examples mentioned by the participants as hateful.

In terms of the reasons behind the spread of hateful speech online, the participants indicated that it could be due to the fact that most of those who spread hateful speech online believe that the Internet is a virtual place, where they feel free to write whatever they like and that no one will find out. Likewise, getting away with such behavior was cited as one of the reasons it was being perpetrated. Moreover, the participants indicated that people are usually bolder online and tend to say things they would not otherwise say in person. They also stated that the environment of some of the digital media platforms was toxic, even encouraging generally decent users to be hostile. Finally, being a huge and wide space, where people of different ages, educational levels, and religions gathered was also mentioned as a reason for this unacceptable behavior.

The participants were asked about their reaction to being exposed to such situations. They were divided into two groups based on their answers: Some indicated that they ignored it, while others believed that they were obliged to react. The first group explained that they did not counteract online hate speech because they believed that those who spread hateful speech were only seeking attention and wanted to gather more followers, retweets, or comments. They also indicated that hateful speech was mostly generated by unknown users with fake accounts. Thus, there was no reason to counteract it. The second group, which tried to counter online hate speech, employed several strategies. Some of the participants indicated that they commented on users' posts to say that their behavior was wrong. Others stated that they reported or blocked such users. Some participants declared that they spread the post to their networks, even using other digital media platforms; thus, they would join them in the process of counteraction. One of the participants even indicated that he spread the post without the need to explain to his network that this was hate speech that should be reported, knowing that they would report it themselves.

## DISCUSSION

The results of phase one demonstrate that most of the sample indicated that they had been victims of online hate speech, that these unpleasant incidents occurred more frequently than being bystanders of hate speech targeting others. On the contrary, very few participants admitted to being online hate preparators compared with the majority indicating that they had never generated online hate speech. This result is remarkable as it shows that compared to other cultures with a high number of online hate speech victims and perpetrators (Obermaier & Schmuck, 2022; Williams, 2019), it seems that Arab users suffer less regarding this form of misconduct. This could be due to the fact that these users belong to a collectivist culture. According to Hofstede (2011), the manifestation of a collectivist culture is a tight, long-term devotion to the people of the group. Such a culture creates close relationships in which everyone has a responsibility to group members. In collectivist societies, offense results in loss of face and humiliation (Obeidat et al., 2012).

The results of phase one revealed the significant positive moderating effect of age on the relationship between hate speech exposure (albeit not victimization) and hate speech perpetration to become stronger as the level of education increased. Previous studies have reported inconsistent results regarding the relationship between age and exposure to or likelihood of being a victim of online hate speech. While a study conducted in four European nations by Hawdon et al. (2017) found that an increasing age increases the probability of exposure to online hate among British and American users of digital media platforms, Kaakinen et al.'s (2018) study, which focused on samples from the same nations, revealed no significant relationship between age as a control variable and the production of hate. It should be noted that these studies considered the role of the age variable alone in relation to hate exposure and production. The present study, however, investigated the role of age as a moderating variable in the relationship between hate speech exposure and perpetration. This study also investigated the interaction between the age variable and other demographic variables and confirmed that the influence of age on the relationship between hate speech exposure and perpetration becomes stronger as the level of education increases. Based on the results of both the present study and previous investigations, it can be argued that it is not quietly appropriate to suggest that users who were exposed to online hate speech as they aged are more likely to produce hateful online content, as completing a higher level of education may cause them to be more likely to realize that such content is harmful when compared with those went through the same unpleasant experience as they aged but completed a lower level of education.

Another notable finding of this study concerned the significant positive moderating effect of education on the relationship between hate speech victimization and hate speech perpetration, which grew stronger as the other demographic variables increased. This finding is in line with the findings of Hall (2018) and Milani et al. (2022), who showed that a higher level of education significantly predicts exposure to online hate content. By contrast, it differs from the finding of Näsi et al. (2021), who revealed that the level of education does not have a significant effect on exposure to online harassment. However, it should be noted that, as with the age variable, the present study did not investigate the role of education alone in the relationship between hate speech victimization and hate speech perpetration. Rather, it investigated the interaction of the education variable with the other demographic variables and confirmed that the influence of education grew stronger as the other demographic variables increased. This finding is interesting and adds to current knowledge through confirming the importance of investigating the interactions between demographic variables rather than focusing on each variable alone when it comes to assessing their moderating roles. Doing so should yield deeper and more informative findings. For example, this study confirmed that exposure to online hate speech could not only trigger more educated users to mimic such behavior among other users but also trigger those who are more educated and older more than those who are younger and less educated.

Through focusing on the role of online disinhibition in the relationship between online hate exposure and perpetration, this study revealed that toxic online disinhibition was the only significant negative moderator of that relationship. Indeed, benign online disinhibition had no impact as a moderator of the relationship between online hate exposure and perpetration. In addition, the impacts of benign and toxic online disinhibition as moderators were not significant in the relationship between online victimization and perpetration. This finding is interesting because it contradicts the finding of Wachs and Wright (2018), who investigated the moderating role of only toxic disinhibition in the relationship between being a bystander to online hate and perpetrating such hate. They found that when bystanders reported greater degrees of toxic online disinhibition, they perpetrated more online hatred. The present finding also contradicts the finding of Harriman et al. (2020), who revealed an association between exposure to online hate messages and benign online disinhibition. Based on Suler’s (2004) definition of online disinhibition as a phenomenon whereby individuals say or do things in cyberspace that they would not say or do during face-to-face interactions because they feel less constrained and more capable of expressing themselves freely, and given the fact that users from Arab cultures appear to employ speech that is more direct in terms of conveying their actual expression or feeling (Anggraeni & Tajuddin, 2018), it could be argued that when the Arab participants in this study were not aware of the extent of their toxic disinhibition on online platforms, they tended to generate more online hate speech when exposed to such speech than when they understood how toxic it was. This is why it seems that the more hateful material the participants were exposed to and the less toxic online disinhibition they experienced, the greater role they played in producing harmful online content.

Furthermore, the impact of high-threshold OCI as a moderator was significant and positive in the relationship between online hate speech exposure and perpetration. On the contrary, the impact of low-threshold OCI as a moderator was positive and significant in the relationship between online hate speech victimization and perpetration. Previous research has found that bystander intervention was positively related to frequent exposure to online hate speech (Obermaier, 2022). However, the current study showed that users who practice high levels of OCI and are exposed to online hate speech perpetrate more hate speech online than those who practice lower levels of intervention. Furthermore, those who practice lower levels of OCI and are themselves victims of online hate speech perpetrate more hate online. It could be argued that high-OCI users generate online hate speech not to attack others for the sake of it, but they use it as a counter speech strategy as they believe that they have a tremendous responsibility to minimize hate speech online. Conversely, it may be that low-OCI users who are victims of online hate speech do not think that these methods are sufficient to lessen hate speech, thereby explaining their hate speech as counter speech.

All in all, the small R2 changes indicate that while these moderating variables had some impact on the relationship, they did not substantially alter or significantly enhance our understanding of the association between hate speech exposure and perpetration. However, it is important to note that even though the R2 changes were small, they should not be disregarded entirely. While they may not account for a substantial amount of additional variance, they still provide valuable insights into the complexity of the relationship and highlight the potential influence of these moderating variables. Future research should explore additional factors that may contribute to a more comprehensive understanding of the relationship between hate speech exposure and perpetration.

The findings of phase two add valuable results and complement the outcomes of phase one in providing a better understanding regarding users' experience with online hate speech and its relationship with online inhibition and OCI. They show that while most of the participants were aware of the spread of hate speech online, some of them considered such acts as counter speech in response to being intimidated online. Another interesting justification mentioned by the participants was that they looked at hate speech as part of freedom of expression. This finding somewhat contradicts that of Llinares and Bellvís (2019). They showed that, in online hate speech, self-censorship was practiced by a large proportion of the sample, that criminal law and content rules–especially the consequences of the law and the surety with which it was implemented–had no direct impact on the choice to convey opinions on the Internet, and that social perception of what others did was crucial.

In terms of the explanations behind such misconduct, the participants mentioned a number of reasons in line with online disinhibition, such as the Internet being a free space, lack of punishment, and anonymity. These reasons substantiated two of Suler's (2004) frequently researched aspects of online engagement: dissociative anonymity and minimization of authority. However, they also mentioned interesting reasons that could contribute to the online disinhibition literature. They indicated that online platforms are toxic environments that encourage even generally decent users to be hostile and that they are huge and wide spaces, where people of different demographics gather. Thus, it should be expected that they would attack and fight each other. This finding is in line with that of Erjavec and Kovačič (2012), that online comments help spread hate speech to the extent that they are the new sites of war between individuals.

Finally, the participants' reported reactions and attitudes toward experiencing online hate speech yielded remarkable results that differ from the classical division of low- and high-threshold OCI (see Kunst et al., 2021). Based on the interviews, the participants could be divided into silent users (i.e., those who have no plan to interfere or react due to the reasons mentioned above) and active users (i.e., those who feel that they are obliged to act to counteract hate speech online and minimize its effect by utilizing a range of strategies). These findings are important as they add to knowledge in the field of OCI.

## CONCLUSIONS, RECOMMENDATIONS, AND LIMITATIONS

This study utilized an explanatory sequential mixed-methods approach involving two phases. More specifically, it employed a quantitative online questionnaire followed by qualitative online vignette interviews to investigate the moderating roles of OCI, online disinhibition, and several demographic variables in the relationship between online hate exposure\victimization and perpetration. These two phases and the subsequent integration of their findings assisted in developing a clearer understanding of the topic under investigation. First, the results of phase one of this study showed that toxic online disinhibition was a negative moderator of the relationship between online hate exposure and perpetration. This finding was complemented by the explanations for such misconduct offered by the participants during phase two, wherein they mentioned reasons such as the Internet being a free and/or toxic space, lack of punishment, and anonymity. Second, the results of phase one also revealed that high-threshold OCI had a positive impact on the relationship between online hate speech exposure and perpetration. Moreover, the impact of low-threshold OCI on the relationship between online hate speech victimization and perpetration was positive. The results of phase two completed these findings by extending the classical division between low- and high-threshold OCI and proposing that online users can be divided into silent users, who have no plan to interfere with or react to online hate speech, and active users, who feel obliged to counteract online hate speech.

When it comes to the roles of the demographic variables, this study was among the first to investigate the interactions between the examined variables as moderators of the relationship between online hate exposure\victimization and perpetration, revealing interesting findings. Indeed, this study confirmed the significant positive moderating effect of age on the relationship between hate speech exposure (albeit not victimization) and hate speech perpetration to become stronger as the level of education increased. It also revealed the significant positive moderating effect of education on the relationship between hate speech victimization and hate speech perpetration, which grew stronger as the other demographic variables increased.

While the current research has several contributions, it also has a number of limitations. For instance, while it utilized a mixed-method approach to deepen its findings, it is a cross-sectional study. Thus, follow-up research is recommended after a considerable number of years. Also recommended is a study focusing on a specific digital media platform and comparing its findings with those reported herein. Besides, it should be noted that the validity of the findings from phase two, which relied on the participants' self-reports of their behaviors and attitudes toward online hate speech, could be affected by factors such as social desirability bias. Thus, the use of a more objective measurement tool for the purpose of data collection is recommended so that future studies increase the reliability of the findings. Finally, a cross-cultural study adapting the same variables as the current study is recommended.

**Funding:** The author received no financial support for the research and/or authorship of this article.

**Ethics declaration:** The author declared that the study was approved by the Ethical Committee at Faculty of Media and Communication, King Abdulaziz University on 12 February 2023. Informed consents were obtained from the participants.

**Declaration of interest:** The author declares no competing interest.

**Data availability:** Data generated or analyzed during this study are available from the author on request.

## REFERENCES

Anggraeni, S. F., & Tajuddin, S. (2018). Expressive speech acts and cultural values in collection of short stories Wahah Al-Asdiqa. *El Harakah* [*The Movement*]*, 20*(1), 99. https://doi.org/10.18860/el.v20i1.4828

Beausoleil, L. E. (2019). Free, hateful, and posted: Rethinking first amendment protection of hate speech in a social media world. *Boston College Law Review, 60*, 2101.

Bormann, M., Tranow, U., Vowe, G., & Ziegele, M. (2022). Incivility as a violation of communication norms–A typology based on normative expectations toward political communication. *Communication Theory, 32*(3), 332-362. https://doi.org/10.1093/ct/qtab018

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities, 18*(3), 297-326. https://doi.org/10.1177/1468796817709846

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semi-structured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research, 42*(3), 294-320. https://doi.org/10.1177/0049124113500475

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech: Systematic review. *Aggression and Violent Behavior, 58*, 101608. https://doi.org/10.1016/j.avb.2021.101608

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

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior, 40*, 108-118. https://doi.org/10.1016/j.avb.2018.05.003

Cheung, C. M., Wong, R. Y. M., & Chan, T. K. (2016). Online disinhibition: Conceptualization, measurement, and relation to aggressive behaviors. In *Proceedings of the 37th International Conference on Information Systems*.

Costello, M., & Hawdon, J. (2018). Who are the online extremists among us? Sociodemographic characteristics, social networking, and online experiences of those who produce online hate materials. *Violence and Gender, 5*(1), 55-60. https://doi.org/10.1089/vio.2017.0048

Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society, 15*(6), 899-920. https://doi.org/10.1080/15205436.2011.619679

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. UNESCO Publishing.

Global Media Insight. (2022). *Social media statistics*. https://www.globalmediainsight.com/

Hall, L. L. (2018). *Race and online hate: Exploring the relationship between race and the likelihood of exposure to hate material online* [Doctoral dissertation, Virginia Tech].

Harriman, N., Shortland, N., Su, M., Cote, T., Testa, M. A., & Savoia, E. (2020). Youth exposure to hate in the online space: An exploratory analysis. *International Journal of Environmental Research and Public Health, 17*(22), 8531. https://doi.org/10.3390/ijerph17228531

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior, 38*(3), 254-266. https://doi.org/10.1080/01639625.2016.1196985

Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture, 2*(1), 2307-2319. https://doi.org/10.9707/2307-0919.1014

Kaakinen, M., Räsänen, P., Näsi, M., Minkkinen, J., Keipi, T., & Oksanen, A. (2018). Social capital and online hate production: A four country survey. *Crime, Law and Social Change, 69*(1), 25-39. https://doi.org/10.1007/s10611-017-9764-5

Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P. (2016). *Online hate and harmful content: Cross-national perspectives*. Taylor & Francis. https://doi.org/10.4324/9781315628370

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

Lapidot-Lefler, N., & Barak, A. (2015). The benign online disinhibition effect: Could situational factors induce self-disclosure and prosocial behaviors? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 9*, 2. https://doi.org/10.5817/CP2015-2-3

Leets, L., & Giles, H. (1999). Harmful speech in intergroup encounters: An organizational framework for communication research. In M. Roloff (Ed.), *Communication yearbook* (pp. 91-137). SAGE. https://doi.org/10.1080/23808985.1999.11678960

Lester, J. N., Cho, Y., & Lochmiller, C. R. (2020). Learning to do qualitative data analysis: A starting point. *Human Resource Development Review, 19*(1), 94-106. https://doi.org/10.1177/1534484320903890

Lim, J. S., & Golan, G. J. (2011). Social media activism in response to the influence of political parody videos on YouTube. *Communication Research, 38*(5), 710-727. https://doi.org/10.1177/0093650211405649

Llinares, F. M., & Bellvís, A. B. G. (2019). Freedom of expression in social media and criminalization of hate speech in Spain: Evolution, impact and empirical analysis of normative compliance and self-censorship. *Spanish Journal of Legislative Studies, 1*, 2019. https://doi.org/10.21134/sjls.v0i1.1837

Meibauer, J. (2013). *Hassrede: Interdisziplinäre Beiträge zu einer aktuellen Diskussion* [*Hate speech: Interdisciplinary contributions to a current discussion*]. Linguistische Untersuchungen [Linguistic Studies].

Milani, R., Caneppele, S., & Burkhardt, C. (2022). Exposure to cyber victimization: Results from a Swiss survey. *Deviant Behavior, 43*(2), 228-240. https://doi.org/10.1080/01639625.2020.1806453

Näsi, M., Danielsson, P., & Kaakinen, M. (2021). Cybercrime victimization and polyvictimization in Finland–Prevalence and risk factors. *European Journal on Criminal Policy and Research, 29*, 283-301. https://doi.org/10.1007/s10610-021-09497-0

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

Obeidat, B. Y., Shannak, R. O., Masa'deh, R. E. M. D. T., & Al-Jarrah, I. (2012). Toward better understanding for Arabian culture: Implications based on Hofstede's cultural model. *European Journal of Social Sciences, 28*(4), 512-522. https://doi.org/10.5296/jmr.v4i4.2160

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

Oksanen, A., Hawdon, J., Holkeri, E., Näsi, M., & Räsänen, P. (2014). Exposure to online hate among young social media users. In *Soul of society: A focus on the lives of children & youth*. Emerald Group Publishing Limited. https://doi.org/10.1108/S1537-466120140000018021

Porten-Cheé, P., Kunst, M., & Emmer, M. (2020). Online civic intervention: A new form of political participation under conditions of a disruptive online discourse. *International Journal of Communication, 14*, 21.

Riedl, M. J., Masullo, G. M., & Whipple, K. N. (2020). The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior, 107*(3), 106262. https://doi.org/10.1016/j.chb.2020.106262

Rizwan, H., Shakeel, M. H., & Karim, A. (2020). Hate-speech and offensive language detection in roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 2512-2522). https://doi.org/10.18653/v1/2020.emnlp-main.197

Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 255-264). https://doi.org/10.1145/3292522.3326032

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society,* 14614448221091185. & Society, 14614448221091185. https://doi.org/10.1177/14614448221091185

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior, 7*(3), 321-326. https://doi.org/10.1089/1094931041291295

Udris, R. (2014). Cyberbullying among high school students in Japan: Development and validation of the online disinhibition scale. *Computers in Human Behavior, 41*, 253-261. https://doi.org/10.1016/j.chb.2014.09.036

Wachs, S., & Wright, M. F. (2018). Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. *International Journal of Environmental Research and Public Health, 15*(9), 2030. https://doi.org/10.3390/ijerph15092030

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

Wachs, S., Gámez-Guadix, M., & Wright, M. F. (2022). Online hate speech victimization and depressive symptoms among adolescents: The protective role of resilience. *Cyberpsychology, Behavior, and Social Networking, 25*(7), 416-423. https://doi.org/10.1089/cyber.2022.0009

Wachs, S., Wright, M. F., & Vazsonyi, A. T. (2019). Understanding the overlap between cyberbullying and cyberhate perpetration: Moderating effects of toxic online disinhibition. *Criminal Behavior and Mental Health, 29*(3), 179-188. https://doi.org/10.1002/cbm.2116

Wilhelm, C., & Joeckel, S. (2019). Gendered morality and backlash effects in online discussions: An experimental study on how users respond to hate speech comments against women and sexual minorities. *Sex Roles, 80*(7), 381-392. https://doi.org/10.1007/s11199-018-0941-5

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research, 47*(6), 921-944. https://doi.org/10.1177/0093650219855330

Williams, M. (2019). *Hatred behind the screens: A report on the rise of online hate speech*. https://orca.cardiff.ac.uk/127085/

Zaharna, R. S. (1995). Understanding cultural preferences of Arab communication patterns. *Public Relations Review, 21*(3), 241-255. https://doi.org/10.1016/0363-8111(95)90024-1

Zhang, Z., & Luo, L. (2018). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. Semantic Web. *Preprint*, 1-21.