Third-Person Effect and Hate Speech Censorship on Facebook

Lei Guo and Brett G. Johnson

Abstract
By recruiting 368 US university students, this study adopted an online posttest-only between-subjects experiment to analyze the impact of several types of hate speech on their attitudes toward hate speech censorship. Results showed that students tended to think the influence of hate speech on others was greater than on themselves. Their perception of such messages’ effect on themselves was a significant indicator of supportive attitudes toward hate speech censorship and of their willingness to flag hateful messages.

Keywords
third-person effect, paternalism, speech freedom support, hate speech, Facebook, censorship

Introduction
Hate speech is a perennial issue on social media (e.g., Citron, 2014). A 2016 study by the Pew Research Center found that a majority of Americans believe the presence of trolling, harassment and hate speech in online discourse will either get worse or not improve (Rainie, Anderson, & Albright, 2017). Some have argued that the 2016 US presidential election emboldened those who would choose to post hateful content on social media due to the fact that caustic rhetoric used by then-candidate Donald Trump during the campaign went largely unchallenged and became normalized (Crandall & White, 2016). Scholars (Citron, 2014; Johnson, 2017) as well as pundits and politicians have questioned whether social media companies have a duty to clean up hate speech on their platforms. Critical to conceiving of the impetus for such a duty is understanding how individual users interpret the impact hate speech has within the context of social networks, as well as users’ willingness to take actions that would eradicate such speech.

Previous research has examined people’s willingness to impose legal restrictions on hate speech in various countries, including the United States, Hungary, and Myanmar, among others (Kvam, 2015; Mudditt, 2015; Reddy, 2002). However, there has been little research conducted on users’ personal attitudes and behaviors in response to hate speech on social media and government regulation. Informed by the third-person effect (TPE) hypothesis, we thus fill the gap by examining such concerns in the context of the three main types of text-based hate speech (racism, sexism, and anti–lesbian, gay, bisexual, transgender [LGBT]) on Facebook, the most popular social network in the world (Lev-On, 2017). By using an online experiment, the main purpose of the study is to determine what role perceived influences of such hate speech have on US students’ willingness to support censorship and willingness to “flag” such content. “Flagging” content involves notifying Facebook of the existence of the content and triggering a process in which human arbiters will examine the speech vis-à-vis certain criteria before deciding to remove it or leave it be (Crawford & Gillespie, 2016). Moreover, the role of paternalism and support for freedoms of expression in shaping individuals’ perception of TPE, support to censor such contents, and willingness to flag these types of hate speech are also tested in the study.

Literature Review and Hypotheses
Hate Speech
Hate speech includes any form of expression, including text, image, and video, and it has been defined in many different ways by many different scholars. A generic definition for the purposes of this study categorizes hate speech as any speech that attacks and attempts to subordinate any group or class of people, typically spoken by a group with a higher level of social power than the targets of the speech (Calvert, 1997; Post, 2009; Tsesis, 2009). The targets of such speech typically

The University of Missouri, USA

Corresponding Author:
Lei Guo, Missouri School of Journalism, The University of Missouri, 181 Gannett Hall, Columbia, MO 65211, USA.
Email: lgfgt@mail.missouri.edu
include racial minorities, women, religious minorities, and those in the LGBT community (Rowbottom, 2009). Such words might not only cause individuals distress, fear, and humiliation but also lead to the spread of violent emotions and even violent crimes in society (Nemes, 2002). Considering the harm of hate speech, many countries and territories have released laws to regulate hate speech. Germany government passed several toughest laws toward hate speech and online hate speech. For example, “NetzDG” requires online platforms to delete “obviously illegal” hate speech within 24hr after receiving the notification; otherwise, they will be fined €50 m (US$55,415,750).

While in the United States, hate speech is legally protected under the First Amendment. In other words, there are no criminal penalties for hate speech, unless the speech falls into one of the nine categories of exception carved out by the counts, which include fighting words, true threats, defamation, and so on (Schauer, 2005). In terms of hate speech on the Internet, Section 230 of the Communications Decency Act (CDA) is even known as a legal shield for online platforms. To protect freedom of speech, it says that “No provider or user of an interactive computer service shall be treated as the information content provider” (47 U.S.C. § 230). It means the online platforms, such as Facebook, Twitter, and YouTube, will not be punished because of users’ posts and online hate speech is protected under Section 230. However, just because such speech is legally allowed does not mean that it is socially acceptable. Indeed, studies have shown that even individuals who support freedom of expression in general harbor less support for hate speech in particular (Yalof & Dautrich, 2002). As the Internet and social media have become more pervasive and have afforded greater anonymity to speakers, the web and social media have become an important space for those willing spread hate speech (Erjavec & Kovačić, 2012; Nemes, 2002). Therefore, online platforms have made efforts to regulate hate speech on the platforms. Taken Facebook as an example, there are over 7,000 human content moderators around the world reviewing users’ posts every day (Koebler & Cox, 2018; Nurik, 2019). In Facebook Community Standards, it clarifies that “Facebook removes hate speech, which includes content that directly attacks people based on their: race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender or gender identity, or serious disabilities or diseases.” However, since this regulation process heavily depends on content reviewers’ personal decision making, controversy has been frequently raised among both Facebook users and reviewers over enforcement of regulation. The state of individuals’ attitudes toward hate speech on social media, viewed via the lens of the TPE, thus becomes a very worthy topic of study.

The Perceptual Component of the TPE

The TPE hypothesis was originally proposed by Davison (1983) to predict that people tend to think the mass media messages’ greatest impact “will not be on ‘me’ or ‘you’ but on ‘them’—the third persons” (p. 3). According to Davison (1983), TPE has two distinctive components: (a) the perceptual component, which deals with an individual’s tendency to believe that the effect of a message on others will be greater than on himself or herself and (b) the behavioral component, which refers to the greater likelihood of individuals to react to a given message in a manner corresponding to self-other perceptual discrepancy.

TPE has been documented in the context of a wide range of messages, including pornography, media violence, political messages, health messages, fashion, controversial products, news about social incidents, the Y2K issue, and the influence of public relations (Lo et al., 2016). Previous research has found that if the media messages are perceived as negative or undesirable (e.g., pornography or violent programs), a TPE is more likely to occur; meanwhile, people are more likely to report that positive messages influence themselves more than others (Buturoiu et al., 2017; Golan & Day, 2008; Gunther, 1995). Several other factors, such as perceived knowledge, news exposure, and social distance, have been shown to either mitigate or amplify the TPE (Lo et al., 2016).

As Perloff (1999) mentioned, TPE “is a reliable and persistent phenomenon that emerges across different content forms and research settings” (p. 369). Although TPE studies have long focused on the perceived influence of traditional mass media, the study of TPE has also been extended to digital media (Kim & Hancock, 2015; Lev-On, 2017). For example, Banning and Sweetser (2007) and Li (2008) confirmed that TPE appeared across blogs and websites. More recently, studies are beginning to explore TPE in the context of Facebook use. By means of a survey with 688 participants, Buturoiu et al. (2017) found that the more users perceived the influence of Facebook content as undesirable, the more they evaluated the effects of such content as being greater on others than on themselves.

In the context of hate speech on Facebook, we expect that participants would perceive such contents to have a greater effect on others than on themselves. The first hypothesis thus can be raised:

HI. Participants who read hate speech on Facebook will perceive a greater effect on the general public than on themselves.

The Behavioral Component of TPE

According to Davison (1983), the self-other perceptual discrepancy might lead people to take remedial actions. This behavioral component of TPE assumed that when people perceived the greater impact of mass communication on others, they would be inclined to take certain actions to mitigate the perceived harmful effects (Bennur, 2008).

Most previous studies have indicated that individuals’ tendency to perceive greater media content influence on others
than on themselves leads to support for censorship or government regulation of media (Golan & Day, 2008; Gunther, 1995; Rojas et al., 1996). Censorship can be defined as a mental condition or action that aims to protect people from the perceived harmful effects of what people read, see, or hear (Dority, 1989). A meta-analysis that conducted by Sun et al. (2008) specifically concluded that undesirable media content such as Internet pornography, violence on television, and rap lyrics would lead to restrictive and corrective behaviors.

Facebook offers an ideal case for studying TPE and the willingness to censor hate speech due to the fact that the social network affords users the ability to react in a censorial manner against instances of hate speech at a relatively low social cost. In particular, Facebook allows users to signal to Facebook which content they wish to have removed by flagging it (Crawford & Gillespie, 2016). In other words, a behavioral component to TPE and willingness to censor is built into Facebook’s architecture. Meanwhile, a traditional behavioral component to TPE—support for government action to censor the social media, in this case, Facebook—still exists in this context. Furthermore, Facebook can voluntarily remove harmful content from its platform as well through their Terms of Service and Community Standards (Carlson & Rousselle, 2020). The Terms of Service, which users must agree to before accessing Facebook, allow Facebook to act similar to the government in its potential to moderate hateful content on its platform (Peters & Johnson, 2016) and offering another lens through which to examine the behavioral component of TPE: support for Facebook voluntarily moderating speech.

Accordingly, in the context of hate speech on Facebook, we anticipate that perceived effect on others, as compared to perceived effects on oneself, is more likely to result in taking actions to censor such messages. Therefore, a series of hypotheses regarding the TPE behavioral component can be raised:

**H2a.** Perceived effects of hate speech on the general public will be more strongly and positively related to willingness to flag hate speech on Facebook than perceived effects on the self.

**H2b.** Perceived effects of hate speech on the general public will be more strongly and positively related to support for Facebook’s content moderation than perceived effects on the self.

**H2c.** Perceived effects of hate speech on the general public will be more strongly and positively related to support for government’s regulation of hate speech on Facebook than perceived effects on the self.

**Paternalism and TPE**

Paternalism is defined as “acting like a parent and limiting others’ rights with good intentions” (Golan & Banning, 2008, p. 211). Several studies have postulated a link between paternalism and TPE. This connection appears logical: paternalism is often associated with a sense that other people are incapable of avoiding negative effects (of media content or other things) and therefore some sort of intervention is the only means to protect them (Golan & Day, 2008; McLeod et al., 1997). Early research regarding the role of paternalism in TPE conducted by McLeod et al. (1997) suggested that paternalism was particularly responsible for explaining the behavioral component of TPE. Indeed, the level of paternalism is considered a significant predictor of support for censorship or government regulations (McLeod et al., 2001; Xu & Gonzenbach, 2008). In addition, McLeod et al. (2001) also identified the relationship between individuals’ paternalistic attitudes and their perceived effects of violent rap lyrics on others. Meanwhile, they also found that their female participants harbored stronger paternalistic attitudes than their male participants, which led them to suggest that the concept could be renamed “maternalism.”

The present study borrows the scale McLeod et al. (2001) used to measure paternalism among participants and assess its role in shaping willingness to censor. Specifically, based on the arguments raised in previous research, we expect that participants’ paternalistic attitudes will be associated with their perceived influences of hate speech on others, as well as their support for censorship and their willingness to flag such content on Facebook. Therefore, the following hypotheses regarding paternalism can be proposed:

**H3a.** Participants’ paternalistic attitudes will be positively related to perceived effects of hate speech on the general public.

**H3b.** Participants’ paternalistic attitudes will be positively related to their willingness to flag hate speech on Facebook.

**H3c.** Participants’ paternalistic attitudes will be positively related to their support for Facebook’s content moderation of hate speech.

**H3d.** Participants’ paternalistic attitudes will be positively related to their support for government’s regulation of hate speech on Facebook.

**Support for Freedom of Speech and TPE**

Freedom of expression is a core value in democracies (Naab, 2012; Post, 2009). As noted above, in the United States, the First Amendment to the US Constitution has been interpreted to allow legal protection for extreme forms of hate speech that would be legally punished in many other parts of the world (Schauer, 2005). However, even with these exceptional legal protections for hate speech, scholars have routinely found that Americans’ tolerance for hate speech has been shown to be low even though their overall support for
freedom of expression, in general, tends to be high (Gibson & Bingham, 1982; Sullivan et al., 1982; Yalof & Dautrich, 2002). That distinction could be evaporating: a 2018 survey by the Knight Foundation found that American university students, especially females, Blacks, and democrats, tend to favor promoting diversity and inclusion on college campuses over protecting the right to freedom of expression (Knight Foundation, 2018).

Previous research has identified that certain personalities, as well as attitudinal and cultural factors, could be significant predictors of individuals’ support for freedom of speech (Downs & Cowan, 2012; Lambe, 2004). However, there is lack of studies about the relationship between support for freedom of expression and TPE and whether individuals’ support for freedom of expression will have an impact on willingness to censor undesirable messages in the context of TPE. We believe studying such relationships is important for two reasons. First, the documented tepid support for freedom of expression among US college students appears to be incongruent with the historical trend of scholars finding a distinction between overall support for freedom of expression and lack of support for hate speech. Second, as social media platforms like Facebook give individuals a built-in means of expressing a willingness to censor speech, it is important to understand whether individuals will use these tools based on their lack of support for freedom of expression.

Despite the absence of evidence from previous research, a linkage between individuals’ support for freedom of speech and TPE was expected in the present study. Because individuals who support free speech would be highly committed to democratic principles, it is possible that their perceived effect of hate speech on others will decrease as they would value individual rights more than group rights (Lambe, 2004). Therefore, our first prediction regarding support of freedom of speech is informed by the above arguments as follows:

\[ H4a. \text{Participants’ support for freedom of expression will be negatively related to the gap between the perceived effects of Facebook hate speech on the general public.} \]

Finally, although previous research has revealed the growing intolerance of hate expression among Americans (Yalof & Dautrich, 2002), considering the development of Internet and freedom of speech, it seems that no consensus has been reached yet on the most desirable method of online hate speech regulation (Nemes, 2002). The existing literature found that measure of individuals’ support for democratic principles was positively related to their political tolerance (Marcus et al., 1995). In other words, when people have a strong belief of democratic principles, they are more likely to assume all forms of speech as equally legitimate or illegitimate (Harell, 2010). Since free speech is one of the most important principles in democracies (Naab, 2012), we can thus assume that those who report a higher degree of support for freedom of expression will be less likely to be willing to censor hate speech on Facebook. Therefore, the following hypotheses can be raised:

\[ H4b. \text{Participants’ support for freedom of expression will be negatively related to their willingness to flag hate speech on Facebook.} \]

\[ H4c. \text{Participants’ support for freedom of expression will be negatively related to their support for Facebook’s content moderation of hate speech.} \]

\[ H4d. \text{Participants’ support for freedom of expression will be negatively related to their support for government’s regulation of hate speech on Facebook.} \]

**Method**

**Sampling**

An online survey experiment was conducted in early December 2017 using a sample of students at a large Midwestern university. Although not a representative sample, a student sample is appropriate for this study because 18- to 34-year-olds are the most active users on Facebook (Statista, 2017), and university students have been proved as an accurate representation of young people in various topics research (Nelson et al., 1997). Furthermore, previous research has found that university students tend to report higher levels of the TPE compared with the general public since they might have more optimistic biases (Rucincki & Salmon, 1990; Salwen & Dupagne, 2001; Willnat, 1996). With the permission of instructors, the online questionnaire was distributed to students in two large undergraduate journalism classes through Qualtrics. To ensure a high response rate, participants were informed that they could receive extra credit for their participation. Finally, 368 students completed the questionnaire.

**Procedures and Experimental Stimuli**

The survey began with questions about participants’ demographics, their media use habits, their attitudes toward freedom of speech, and level of paternalism. Next, we created four different conditions and embedded in the survey. Participants were randomly assigned to one of four conditions. In the control condition, subjects were asked to estimate the effect of exposure to the rather common and innocuous Facebook message “Thank goodness it’s Friday!” The remaining three experimental conditions were designed to resemble hateful responses to major political issues involving race, gender, and the LGBT community that may commonly be found on Facebook feeds. Participants in these three groups were randomly assigned to a condition in which they were exposed to one of the following types of hate speech: racist speech, sexist
speech, or anti-LGBT speech. The message “Don’t want any Mexicans or Muslims in my country. #ProtectOurHeritage #WhiteLivesMatter” was used as the racist speech condition. The message “Stop whining about free birth control and get back in the kitchen like a good woman should” was used as the sexist speech condition. The anti-LGBT condition was presented via the message “Homosexuality is an abomination. As long as we allow sodomites to marry, America will continue down its road to Hell.”

The reason that participants were given only one category of hate speech as a stimulus was so that the possible effects of seeing one type of speech did not interfere with seeing other types of speech. In other words, the study sought to avoid a priming effect of one type of speech on reactions to other types of speech (Iyengar et al., 1982; Smith et al., 2003).

Participants were asked to indicate their perceived effects of such speech on self and on the general public. They were also asked to indicate their support for Facebook taking censorship actions against the speech, their support for government regulations of such speech, and their willingness to flag such speech. In this way, the resulting data can be analyzed on a between-subjects basis in terms of the different types of Facebook hate speech. They can also be examined on a within-subjects basis to determine whether the self-others perceptual gap is affected by using different hate speech stimuli.

Measurement

Perceived Effects of Social Media Hate Speech on the Self and Others. In the racist speech condition, participants were asked to indicate to what extent the speech influenced them in terms of (a) attitude toward immigrants, (b) attitude toward immigration policy, (c) attitude toward non-White people, and (d) attitude toward anti-discrimination policies. Participants were asked to the extent to which sexist speech influenced them in terms of (a) attitude toward men, (b) attitude toward women, (c) attitude toward sexism, and (d) attitude toward anti-discrimination policy. Participants were asked to indicate to what extent anti-LGBT speech influenced them in terms of (a) attitude toward LGBT people, (b) attitude toward homophobia, (c) attitude toward legalization of same-sex marriage, and (d) attitude toward anti-discrimination policies. Finally, in the control condition, participants were asked to indicate to what extent the speech influenced them in terms of (a) attitude toward Fridays, (b) attitude toward standard work weeks, (c) attitude toward people who work 9–5 jobs, (d) attitude toward people who complain about the work week, and (e) attitude toward labor policies. The response categories ranged on a 7-point Likert-type scale, where 1 = “no influence at all” and 7 = “a great deal of influence.” The four “self” items were averaged to create a measure of “perceived racist hate speech effects on the self” (= .95), “perceived sexist hate speech effects on the self” (= .85), “perceived anti-LGBT hate speech effects on the self” (= .91), and “perceived non-hate speech effects on the self” (= .95). The measure of the perceived effect of such messages on the general public consisted of four parallel items of each condition (replacing “you” with “general public”), including “perceived racist hate speech effects on the general public” (= .95), “perceived sexist hate speech effects on general public” (= .87), “perceived anti-LGBT hate speech effects on the general public” (= .94), and “perceived non-hate speech effects on the general public” (= .95).

Paternalism. Following McLeod et al.’s (2001) research, participants were asked to indicate their level of agreement with five statements: (a) sometimes it is necessary to protect people from doing harm to themselves; (b) it is important for the government to take steps to ensure the well-being of citizens; (c) if people are unable to help themselves, it is the responsibility of others to help them; (d) some people are better than others at recognizing harmful influences; and (e) just because people are unable to help themselves, it does not mean the government should step in and try to help them. The 7-point response categories ranged from 1 = “completely disagree” to 7 = “completely agree.” The last item was deleted because of low reliability. The other four items were combined to create an overall “paternalism” score (\(M = 5.62\), standard deviation \([SD] = 0.82\), \(= .71\)).

Support for Freedom of Speech. Participants also reported the extent of their support for speech freedom, using 1 = “strongly disagree” to 7 = “strongly agree.” The items were based on previous studies that examined support for freedom of expression (Sullivan et al., 1982; Yalof & Dautrich, 2002) and included the following statements: (a) in general, I support the First Amendment, (b) freedom of expression is essential to democracy, (c) democracy works best when citizens communicate in an unregulated marketplace of ideas, and (d) even extreme viewpoints deserve to be voiced in society. The means of these four items were computed as “support for freedom of speech” (\(M = 6.06, SD = 0.77\), \(= .76\)).

Support for Hate Speech Censorship. We developed six items to measure the extent of participants’ support censoring hate speech on Facebook. Participants were asked to indicate their agreement on a 7-point scale from 1 = “strongly disagree” to 7 = “strongly agree.” The statements were as follows: (a) I support to Facebook taking action to censor hate speech; (b) I support Facebook setting up more functions to let users block hate speech; (c) I support Facebook setting up more functions to let users report hate speech to Facebook; (d) I support Facebook closing the accounts of individuals or groups who post hate speech; (e) I support the government passing a law to clean up hate speech; and (f) I support government passing a law to punish individual Facebook users for their hate speech. The first four items were combined to a measure of “support for Facebook content moderation” (\(M = 5.14, SD = 1.31\), \(= .88\)), and last two items were computed as
Willingness to Flag Hate Speech. We developed four items to measure willingness to flag hate speech on Facebook. Participants were asked about their attitudes toward the statement: I am willing to flag this message. The 7-point scale ranged from 1 = “extremely unlikely” to 7 = “extremely likely.” The means of the 7-point scales were labeled as “willingness to flag the speech” (\(M=3.12, SD=1.74, \rho=.76\)).

Control Variables. Participants were asked about their gender (0 = male, 1 = female), grade (1 = freshman, 2 = sophomore, 3 = junior, 4 = senior), Facebook exposure (1 = less than 1 hr/day, 2 = 1–2 hr/day, 3 = 3–4 hr/day, 4 = more than 4 hr/day), and political affiliation (1 = very liberal, 7 = very conservative). These variables were used as controls in analyses because previous studies indicated that they were related to TPE (Gunther, 1995; Lee, 2009; Lo & Chang, 2006).

Table 1. Partial Correlation Matrix of Study Variables (\(N=368\)).

| Variables | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|-----------|-----|-----|-----|-----|-----|-----|-----|
| 1 Facebook exposure | .09 | | | | | | |
| 2 Support for freedom of expression | .11* | .16** | | | | | |
| 3 Paternalism | | | | | | | |
| 4 Perceived effects on self | -.05 | -.15** | .07 | | | | |
| 5 Perceived effects on the general public | -.04 | -.03 | .21*** | .59*** | | | |
| 6 Support for Facebook content moderation | .06 | -.17** | .41*** | .18*** | .15*** | | |
| 7 Support for government regulation | -.01 | -.28*** | .24*** | .26*** | .20*** | .64*** | |
| 8 Willingness to flag hate speech on Facebook | -.19*** | -.06 | .10 | .35*** | .44** | .14** | .17** |

*p < .05; **p < .01; and ***p < .001.

“support for government regulation” (\(M=3.54, SD=1.85, \rho=.85\)).

“Willingness to Flag Hate Speech.” We developed four items to measure willingness to flag hate speech on Facebook. Participants were asked about their attitudes toward the statement: I am willing to flag this message. The 7-point scale ranged from 1 = “extremely unlikely” to 7 = “extremely likely.” The means of the 7-point scales were labeled as “willingness to flag the speech” (\(M=3.12, SD=1.74, \rho=.76\)).

Control Variables. Participants were asked about their gender (0 = male, 1 = female), grade (1 = freshman, 2 = sophomore, 3 = junior, 4 = senior), Facebook exposure (1 = less than 1 hr/day, 2 = 1–2 hr/day, 3 = 3–4 hr/day, 4 = more than 4 hr/day), and political affiliation (1 = very liberal, 7 = very conservative). These variables were used as controls in analyses because previous studies indicated that they were related to TPE (Gunther, 1995; Lee, 2009; Lo & Chang, 2006).

Results

Of the sample, 267 (72.6%) were women, 98 (26.6%) were men, and three were transgender (0.8%). Nearly all of the participants (96.7%) were between 18 and 24 years old. Most of the participants in the sample were White (76.6%), followed by Asian (13.0%), Black or African American (4.9%), Latino or Hispanic (2.4%), Native Hawaiian or Pacific Islander (1.1%), and American Indian or Alaska Nation (0.3%). Finally, for their political affiliation, 70 (19.0%) participants reported that they were independent or neutral, 197 (53.5%) were liberal or pro-liberal, and 101 (27.5%) were conservative or pro-conservative.

In order to test the proposed hypotheses and research questions, a partial correlation matrix was built after controlling for gender, race, and party affiliation (Table 1). It showed that perceived effects on self and on the general public were significantly correlated with support for Facebook content moderation, government regulation, and willingness to flag hate speech. Meanwhile, support for speech freedom was negatively associated with support for Facebook and government’s censorial actions, and paternalism was highly correlated with perceived effects on the general public, support for content moderation by Facebook, and willingness to flag such speech, except with perceived effects on the self.

H1 predicted that participants who read hate speech would perceive greater effects on the general public than on themselves. After controlling for gender, race, political affiliation, Facebook exposure, support to speech freedom, and paternalism, a repeated-measures analysis was conducted with between-factor as four different conditions and within-factors as perceived effect on self and on the general public. The results demonstrated a significant mean difference in perceived effects between the four experimental groups (\(F(3, 358)=23.19, p<.001, \eta^2=.16\)). A follow-up pairwise comparison indicated that participants in the control group (\(M_{Adj}=2.68, SE=0.15\)) reported low level of perceived effects than participants in the racist (\(M_{Adj}=3.64, SE=0.16, p<.001\)), sexist (\(M_{Adj}=4.37, SE=0.14, p<.001\)), and anti-LGBT (\(M_{Adj}=3.64, SE=0.13\)) hate speech groups.

The difference between participants’ perceived effect of hate speech on themselves and on the general public was further examined. However, the data indicated at least in one group, there was no difference between these two effect perceptions (\(F(1, 358)=1.01, p=.32, \eta^2=.003\)). Again, a follow-up pairwise comparison provided support for the baseline TPE for the effects of the comparison groups except for the control group. Specifically, as Table 2 shows, in the racism condition, participants perceived this message to have a greater effect on the general public (\(M_{Adj}=4.28, SE=0.16\)) than on themselves (\(M_{Adj}=3.01, SE=0.19, p<.001\)). Similarly, they tended to agree that the general public (\(M_{Adj}=4.74, SE=0.15\)) would be more affected by sexist speech than themselves (\(M_{Adj}=4.00, SE=0.17, p<.001\)). The same phenomenon is also generated under the anti-LGBT hate speech condition; the results showed that the participants believed that compared with the general public (\(M_{Adj}=4.38, SE=0.14\)), they were less likely to be influenced by such a message (\(M_{Adj}=2.90, SE=0.16, p<.001\)). Also, as expected, there was no difference between participants’ perceived effect on self (\(M_{Adj}=2.58, SE=0.18\)) and
on the general public (MAdj = 2.79, SE = 0.15, p = .17) in the control group. Therefore, H1 was statistically supported by the data.

H3a predicted that participants’ paternalistic attitudes would be positively related to perceived effects of hate speech on the general public. According to the repeated-measures analysis, support for freedom of speech (F(1, 358) = 4.31, p < .001, \( \eta^2 = .012 \)) had a main effect with perceived effects of Facebook hate speech. To test the specific relationship between these two variables, a series of hierarchical regression analyses was performed. As Table 3 indicates, in the sexist hate speech context, paternalism was highly and positively related to participants’ perception of the impacts of such hate speech on the general public (\( \beta = .25, p = .04 \)). In the anti-LGBT hate speech context, the paternalistic mental state was found as a significant predictor of perceived effects of such content on self (\( \beta = .26, p = .017 \)) and the general public as well (\( \beta = .53, p < .001 \)). Yet, in the racist hate speech condition, paternalism failed to predict perceived effects on self (\( \beta = .03, p = .81 \)) and general public (\( \beta = .13, p = .32 \)). Therefore, H3a was partially supported.

H4a predicted that participants’ support for freedom of speech would predict their perceived effects of hate speech on people in general. The former repeated-measures analysis indicated that there was a relationship between support for freedom of speech (F(1, 358) = 13.81, p < .001, \( \eta^2 = .037 \)) and perceived effects of Facebook hate speech. A further series of hierarchical regression analyses were performed to confirm the specific relationship between these two variables. Table 3 shows that under the three types of hate speech, neither perceived effects on self nor perceived effects on the general public could be predicted by support for freedom of speech. Therefore, H4a was rejected by the data. This might indicate that the relationship between support for freedom of speech and the perceived effect of hate speech on the general public might not be a linear relationship or might be mediated by other factors. Again, as expected, under the non-hate speech condition, the level of participants’ paternalism and support for freedom speech wound not affect their perceived of such message on the general public.

To address willingness to flag hate speech on Facebook, a series of hierarchical regression analyses were performed to determine factors that contributed to this behavior. Table 2 presents the mean estimates of perceived effects of hate speech on Facebook on self and the general public (N = 368).

### Table 2. Mean Estimates of Perceived Effects of Hate Speech on Facebook on Self and the General Public (N = 368).

| Perceived effect       | Control | Racist hate speech | Sextist hate speech | Anti-LGBT hate speech |
|------------------------|---------|--------------------|---------------------|-----------------------|
|                        | MAdj    | SE     | N   | MAdj | SE     | N   | MAdj | SE     | N   | MAdj | SE     | N   |
| On self                | 2.58    | 0.18   | 88  | 3.01 | 0.19   | 77  | 4.00 | 0.17   | 95  | 2.90 | 0.16   | 108 |
| On general public      | 2.79    | 0.15   | 88  | 4.28 | 0.16   | 77  | 4.74 | 0.15   | 95  | 4.38 | 0.14   | 108 |

LGBT: lesbian, gay, bisexual, transgender; SE: standard error.

### Table 3. Hierarchical Regression Analysis Predicting Perceived Effects on Self and Perceived Effects on the General Public (N = 368).

| Independent variables | Perceived effects of racist hate speech | Perceived effects of sexist hate speech | Perceived effects of anti-LGBT hate speech | Control |
|-----------------------|----------------------------------------|----------------------------------------|-------------------------------------------|---------|
| Block 1: demo         |                                        |                                        |                                           |         |
| Gender                | .11                                    | .19                                    | .16                                      | .01     |
| Race                  | –.11                                   | .05                                    | .26*                                     | .04     |
| Political affiliation | .08                                    | .07                                    | –.14                                     | –.04    |
| Incremental \( \Delta R^2 \) | –1.9%                                 | –0.2%                                 | 8.1%                                     | –2.3%   |
| Block 2: media exposure |                                      |                                        |                                           |         |
| Facebook exposure     | .17                                    | .06                                    | –.03                                     | –.03    |
| Incremental \( \Delta R^2 \) | 1.1%                                  | –1.3%                                 | –1.1%                                    | –1.3%   |
| Block 3: speech freedom |                                      |                                        |                                           |         |
| Support for freedom of speech | –.07                                  | .06                                    | –.10                                     | .02     |
| Incremental \( \Delta R^2 \) | –1.1%                                  | –1.3%                                 | –0.7%                                    | –1.1%   |
| Block 4: paternalism  |                                        |                                        |                                           |         |
| Paternalism           | .03                                    | .13                                    | .12                                      | .25*    |
| Incremental \( \Delta R^2 \) | –1.6%                                  | 0                                     | 0.1%                                     | 4.6%    |
| Total \( \Delta R^2 \) | –3.5%                                  | –2.8%                                 | 6.4%                                     | –0.1%   |

LGBT: lesbian, gay, bisexual, transgender.

1 = perceived effects on self, 2 = perceived effects on general public.

* \( p \leq .05 \); ** \( p \leq .01 \); and *** \( p \leq .001 \).
find out the most significant predictor of behavioral intention outcomes. As shown in Table 4, when focusing on the specific conditions, neither participants’ paternalistic attitudes nor support for freedom of expression had a significant relationship with participants’ willingness to flag hate speech. In addition, only the perceived effect of sexist hate speech on self ($β = .26, p < .001$) could predict participants’ consequent censorial behavior. Therefore, H3b and H4b were rejected by the data, and H2a was only partially supported. The same analysis was conducted to clarify the significant predictors of support for Facebook’s content moderation of hate speech. The results showed that in the racist hate speech context, paternalism ($β = .36, p = .004$) and perceived effects on self ($β = .25, p = .048$) could lead to support for content moderation of hate speech on Facebook, while paternalism ($β = .26, p = .049$) was a significant indicator of support for government regulation of racist speech. Within the sexist hate speech condition, gender ($β = .32, p = .002$) and perceived effects on the general public ($β = .34, p = .002$) would predict participants’ supportive attitudes toward Facebook censorship, while the lack of support for freedom of expression ($β = .27, p = .014$) and perceived effects on self ($β = .25, p = .045$) might lead to more support for government regulation of sexist hate speech. For anti-LGBT hate speech, surprisingly, the greater the participants’ exposure to Facebook ($β = .24, p = .007$), the more likely that they would not support for Facebook acting to moderate this type of hate speech. In addition, the lack of support for freedom of expression ($β = .28, p = .002$) and the high level of paternalism ($β = .56, p < .001$) were the most important predictors of support for Facebook content moderation. Meanwhile, lack of support for freedom of expression ($β = .34, p < .001$), paternalistic attitude ($β = .28, p = .009$), and perceived effects of anti-LGBT hate speech on self ($β = .24, p = .024$) were highly related to support for government regulation. Therefore, because perceived effects of sexist hate speech on the general public were only related to stronger supportive attitudes toward Facebook moderating sexist speech, the data partially supported H2b and H2c. H3c and H3d were supported in the racist and anti-LGBT but not the sexist context, meaning H3c and H3d were partially supported. Similarly, as support for freedom of expression could not predict supportive attitudes toward censorial actions by either Facebook or the government in the racist contexts, H4c and H4d were partially supported by the data.

### Discussion

TPE is commonly used in examining people’s supportive attitude of censorship and related behaviors or behavioral intentions (Davison, 1983; Lo et al., 2016). The debates surrounding the censorship of hate speech on Facebook have

---

### Table 4. Hierarchical Regression Analysis Predicting Support for Facebook Content Moderation, Government Regulation, and Willingness to Flag Hate Speech on Facebook (N = 368)

| Independent variables | Racist hate speech | Sexist hate speech | Anti-LGBT hate speech | Control |
|-----------------------|--------------------|--------------------|-----------------------|---------|
| 1                     | 2                  | 3                  | 1                     | 2       | 3       |
| Gender                | .19                | .09                | .07                   | .32**   | .04     | .23*    | .03     | .02     | .01     | .19     | .19     | .01     |
| Race                  | −.04               | −.06               | −.06                  | .09     | −.19    | −.11    | −.12    | −.11    | −.16    | .12     | .18     | −.10    |
| Political affiliation | −.09               | .11                | −.25*                | −.18    | −.04    | .13     | −.08    | .06     | −.06    | −.17    | −.23*   | .05     |
| Incremental $ΔR^2$    | 5.0%               | −1.6%              | 4.9%                  | 16.2%   | 0%      | 7.4%    | −2.1%   | 0.2%    | −0.7%   | 14.1%   | 12.0%   | 8.0%     |
| Block 2: media exposure | Facebook exposure | .06                | −.00                 | −.04    | .16     | .02     | −.00    | −.24**  | −.13    | −.30**  | .01     | .02     | −.23*   |
| Incremental $ΔR^2$    | −1.2%              | −1.5%              | −1.3%                 | 0.2%    | −1.4%   | −1.2%   | 6.4%    | 2.9%    | 9.9%    | −1.1%   | −1.1%   | 6.7%     |
| Block 3: speech freedom | Support for freedom of speech | −.11               | −.05                 | .04     | −.18    | −.27*   | −.12    | −.28**  | −.34*** | −.01    | −.27**  | −.40*** | −.17    |
| Incremental $ΔR^2$    | −0.1%              | −0.7%              | −1.3%                 | 1.8%    | 6.7%    | 0.6%    | 2.8%    | 7.8%    | −1.0%   | 2.1%    | 11.3%   | 4.0%     |
| Block 4: paternalism  | Paternalism        | .36**              | .26*                 | .21     | .01     | .03     | .09     | .56***  | .28**   | .12     | .49***  | .28*    | −1.9%    |
| Incremental $ΔR^2$    | 9.5%               | 5.0%               | 3.7%                  | 0       | −0.2%   | 0.4%    | 28.9%   | 14.0%   | 1.9%    | 17.3%   | 4.8%    | 1.5%     |
| Block 5: TPE          | Effects on self    | .25*               | .25                  | .16     | .13     | .25*    | .39**   | .14     | .24*    | .06     | .41     | .25     | .10      |
| Effects on the public | −.16               | 0                  | .16                  | .34**   | .21     | −.01    | −.08    | .09     | .07     | −.47*   | −.21    | .31      |
| Incremental $ΔR^2$    | 3.2%               | 3.3%               | 4.2%                  | 14.7%   | 11.5%   | 11.4%   | 0       | 6.3%    | −1.2%   | 1.9%    | 0.9%    | 11.9%    |
| Total $ΔR^2$          | 16.4%              | 3.9%               | 10.2%                 | 32.9%   | 16.6%   | 18.6%   | 36.0%   | 31.2%   | 8.9%    | 22.2%   | 26.1%   | 32.1%    |

**LGBT**: lesbian, gay, bisexual, transgender; TPE: third-person effect.  
1 = Facebook censorship, 2 = government regulation, 3 = willingness to flag.  
*p < .05; **p < .01; and ***p < .001.
surfaced more and more in recent years due to the increasing essentiality of social media in political discourse and the concomitant rise of political extremism. The present research extended the TPE research to the context of managing and censoring hate speech on Facebook.

The results produced strong support for the TPE hypothesis. As expected, participants perceived the Facebook racist, sexist, and anti-LGBT hate speech conditions to have a greater influence on others than on themselves. Given people’s intention that protect others from undesirable messages, the results of the study further revealed that participants’ paternalistic attitude was a significant predictor of perceived greater effects of Facebook hate speech on the general public instead of themselves. However, such predictive power of paternalism failed to indicate the perceived effect on self and others in the racist hate speech context. In other words, even participants who reported high levels of paternalism would not perceive any effects of racist hate speech on self and on the general public. It might indicate that participants might not take such hate speech on Facebook as seriously as sexist or anti-LGBT messages. In addition, as racist hate speech is pervasive on Facebook (Siapera et al., 2018), the other potential reason might be because people have become accustomed to such messages and failed to identify such messages as racist hate speech. Future research could thus further explore the mediation effect between paternalism and effects perception or include more potential independent variables in predicting perceived effects on oneself and others, especially under the racist hate speech condition. Surprisingly, support for freedom of speech was not found to be significantly related to the perceived effects of Facebook hate speech on oneself and on the general public. This finding suggests that participants interpreted the effects of exceptional freedom of expression (which includes legal protections for hate speech) was related neither to themselves nor to others.

Another goal of the study was to explore how the perceived effects of Facebook hate speech affected participants’ consequent censorial behavior. The results showed that when females perceived sexist speech on Facebook having more effects on the general public, they would be more willing to support Facebook and government taking actions to moderate and censor sexist hate speech. On the other hand, the perceived effect of such speech on themselves was found to have an effect on female participants’ intention to flag it on Facebook. Gender-based hate speech is prevalent in cyberspace, and women and girls are usually targeted and by such hate speech (Chetty & Alathur, 2018). These features of sexist hate speech validate our finding that US university students, especially female students, are more willing to regulate this type of hate speech on Facebook.

Surprisingly, contrary to our expectation, in racist and anti-LGBT hate speech contexts, the perceived effect on the general public failed to predict participants’ consequent censorial activities. However, the perceived effect on self could lead to support for Facebook’s content moderation of racist hate speech and the government’s regulation of anti-LGBT hate speech, respectively. Informed by previous research, the findings indicate that instead of supporting regulation, people might think that educating others is a more reasonable solution when they perceived racist and anti-LGBT hate speech to have effects on others (Jang & Kim, 2018). In other words, the current study revealed that the perceived effect of Facebook hate speech on oneself was a better predictor of US students’ supportive attitude toward censorial behaviors. This is a theoretically significant finding of this study since it helps advance TPE research by showing that perceived effects on self and others play different roles in triggering censorial behavior. It further confirms that the TPE phenomenon in politically significant cases might be different from it in other contexts (Wei et al., 2019). A possible explanation may be that these types of hate speech are usually highly related to political topics in the United States, making these results consistent with previous research (e.g., Golan & Day, 2008) that has found that the perceived effect of political messages on self prompted people to participate in political action.

However, interestingly, the results also revealed that although Facebook provides users autonomy to express their willingness to censor speech, participants were less likely to flag racist and anti-LGBT hate speech on Facebook even when they reported high level of paternalism, support for freedom of expression, and perceive high impact of such speech on themselves and the public. This pattern of students’ censorial behavior is consistent with previous research which found that US university students’ attitudes toward political activities are largely gripped by apathy and cynicism (Harvard IOP Youth Poll, 2019; Kohnle, 2013). It is also worth noting that except in the racist hate speech context, those with a higher level of supportive attitudes toward free speech were less likely to support content moderating sexist and anti-LGBT hate speech on Facebook. The result could be explained by the root of First Amendment protection or free-speech rights in the US society.

Several limitations of this study should be recognized. First, the imbalance of participants’ demographic profile of the sample may serve as a limitation to the accuracy of the findings. To be specific, of the 368 valid respondents, over 72% were female, while may not represent general Facebook users in the United States. Because based on Facebook statistics, as of February 2020, users are 54% female and 46% male. The respondents were also more prone to be white compared with the entire Facebook users in the United States. Moreover, the latest report of free expression on US college students found that female White students were more likely to agree that hate speech should be protected by the First Amendment. In other words, they are somewhat less likely than other college students to express support for hate speech censorship (Knight Foundation, 2019). Future research could consider using a more representative sample in order to better understand social media users’ censorial attitudes toward online hate speech. Furthermore, we found
that most of the TPE variables in this study failed to predict the behavioral component of TPE. In addition to the sampling bias, one plausible explanation of this limitation is that the participants could not distinguish between the comparison groups devised in the current study. According to Lo (2000), demographic characteristics like educational level and age could affect the self-other gap, which in turn could indirectly influence the behavioral outcomes of TPE. Future studies should examine other comparison groups (e.g., other Facebook users) to assess the predictive power of the self-other gap to TPE outcomes. Finally, although we were confident that our stimuli reflected common examples of hate speech found on Facebook, they by no means reflected the most extreme examples of hate speech out there, leading the results of this study to be relatively small. Previous hate speech studies suggested that targeted groups were more likely than the general public to be affected by hate speech. Therefore, instead of asking respondents perceived the effects of hate speech on the general public, the following studies could differentiate between attitudes of respondents toward themselves and the targeted groups. It is also certainly possible that scholars could find more robust results by using more extreme examples of different forms of hate speech as stimuli.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Lei Guo https://orcid.org/0000-0001-8480-8688

References
Banning, S. A., & Sweetser, K. D. (2007). How much do they think it affects them and whom do they believe? Comparing the third–person effect and credibility of blogs and traditional media. Communication Quarterly, 55(4), 451–466. https://doi.org/10.1080/01463370701665114

Bennur, Z. (2008). Behavioral consequences of the third-person effect on Turkish voters [Doctoral thesis, Middle East Technical University].

Buturoiu, R., Durach, F., Udrea, G., & Corbu, N. (2017). Third-person perception and its predictors in the age of Facebook. Journal of Media Research, 10(2), 18–36. https://doi.org/10.24193/jmr.28.2

Calvert, C. (1997). Hate speech and its harms: A communication theory perspective. Journal of Communication, 47(1), 4–19. https://doi.org/10.1111/j.1460-2466.1997.tb02690.x

Carlson, C. R., & Rousselle, H. (2020). Report and repeat: Investigating Facebook’s hate speech removal process. First Monday, 25. https://journals.uci.edu/ojs/index.php/fm/article/view/10288/8327

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

Citron, D. K. (2014). Hate crimes in cyberspace. Harvard University Press.

Crandall, C. S., & White, M. H., II. (2016, November 17). Trump and the social psychology of prejudice. Undark. https://undark.org/article/trump-social-psychology-prejudice-unleashed/

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. New Media & Society, 18(3), 410–428. https://doi.org/10.1177/146144481543163

Davison, W. P. (1983). The third-person effect in communication. Public Opinion Quarterly, 47(1), 1–15. https://doi.org/10.1086/268763

Dority, B. (1989). Feminist moralism, “pornography,” and censorship. The Humanist, 49(6), 8–9.

Downs, D. M., & Cowan, G. (2012). Predicting the importance of freedom of speech and the perceived harm of hate speech. Journal of Applied Social Psychology, 42(6), 1353–1375. https://doi.org/10.1111/j.1559-1816.2012.00902.x

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

Gibson, J. L., & Bingham, R. D. (1982). On the conceptualization and measurement of political tolerance. The American Political Science Review, 76(3), 603–620. https://doi.org/10.1017/S0003055400188434

Golan, G. J., & Banning, S. A. (2008). Exploring a link between the third-person effect and the theory of reasoned action: Beneficial ads and social expectations. American Behavioral Scientist, 52(2), 208–224. https://doi.org/10.1177/000276408321352

Golan, G. J., & Day, A. G. (2008). The third-person effect and its behavioral consequences: A new trend in twenty-five-year history of third-person effect research. Mass Communication and Society, 11(4), 536–556. https://doi.org/10.1080/15205430802368621

Gunther, A. C. (1995). Overrating the X-rating: The third-person perception and support for censorship of pornography. Journal of Communication, 45(1), 27–38. https://doi.org/10.1111/j.1460-2466.1995.tb00712.x

Harel, A. (2010). The limits of tolerance in diverse societies: Hate speech and political tolerance norms among youth. Canadian Journal of Political Science, 43(2), 407–432. https://doi.org/10.1017/S0008423910000107

Harvard IOP Youth Poll. (2019). New Harvard Institute of Politics survey of young Americans find correlation between anxiety and views related to direction of the country. https://iop.harvard.edu/about/newsletter-press-release/spring-2019-harvard-iop-youth-poll-results

Iyengar, S., Peters, M. D., & Kinder, D. R. (1982). Experimental demonstrations of the “not-so-minimal” consequences of television news programs. The American Political Science Review, 76(4), 848–858. doi.org/10.1017/S000305540018966X

Jang, S. M., & Kim, J. K. (2018). Third person effects of fake news: Fake news regulation and media literacy interventions. Computers in Human Behavior, 80, 295–302. https://doi.org/10.1016/j.chb.2017.11.034

Johnson, B. G. (2017). Speech, harm, and the duties of digital intermediaries: Conceptualizing platform ethics. Journal of Media
Statista. (2017). Distribution of Facebook users worldwide as of January 2017, by age and gender. https://www.statista.com/statistics/376128/facebook-global-user-age-distribution/

Sullivan, J. L., Pireson, J., & Marcus, G. E. (1982). Political tolerance and American democracy. The University of Chicago Press.

Sun, Y., Pan, Z., & Shen, L. (2008). Understanding the third-person perception: Evidence from a meta-analysis. *Journal of Communication, 58*(2), 280–300. https://doi.org/10.1111/j.1460-2466.2008.00385.x

Tsesis, A. (2009). Dignity and speech: The regulation of hate speech in a democracy. *Wake Forest Law Review, 44*(2), 497–532.

Wei, R., Lo, V. H., & Zhu, Y. (2019). Need for orientation and third-person effects of the televised debates in the 2016 U.S. presidential election. *Mass Communication and Society, 22*, 565–583. https://doi.org/10.1080/15205436.2019.1601227

Willnat, L. (1996). Mass media and political outspokenness in Hong Kong: Linking the third-person effect and the spiral of silence. *International Journal of Public Opinion Research, 8*, 187–212. https://doi.org/10.1093/ijpor/8.2.187

Xu, J., & Gonzenbach, W. J. (2008). Does a perceptual discrepancy lead to action? A meta-analysis of the behavioral component of the third-person effect. *International Journal of Public Opinion Research, 20*(3), 375–385. https://doi.org/10.1093/ijpor/edn031

Yalof, D. A., & Dautrich, K. (2002). *The First Amendment and the media in the court of public opinion*. Cambridge University Press.

**Author Biographies**

Lei Guo is a doctoral student at the University of Missouri. Her research interests include journalists’ vocational psychology, media sociology, and media effects.

Brett G. Johnson (PhD, University of Minnesota) is an assistant professor of journalism studies at the University of Missouri. His research interests include the intersection of media law, media ethics, and media sociology.