# Responding to Online Toxicity: Which Strategies Make Others Feel Freer to Contribute, Believe That Toxicity Will Decrease, and Believe that Justice Has Been Restored?

Alison I. Young Reusser1, Kristian M. Veit2, Elizabeth A. Gassin2, and Jonathan P. Case1

1 Houghton University

2 Olivet Nazarene University

# Author Note

Alison Young Reusser [](https://orcid.org/0000-0001-6905-1832) <https://orcid.org/0000-0003-4822-6335>

Pilot data and additional online materials are or will be openly available at the project’s Open Science Framework page (<https://osf.io/6dwjx/?view_only=2b45b35cf37e46e5818a40bf79fc981d>). We have no conflicts of interest to disclose. This research has been funded by a Networking Grant generously provided by the Council of Christian Colleges and Universities.

Correspondence concerning this article should be addressed to Alison Young Reusser, Houghton University, One Willard Avenue, Houghton, NY 14744. Email: [alison.youngreusser@houghton.edu](mailto:alison.youngreusser@houghton.edu)

# Abstract

When we encounter toxic comments online, how might individual efforts to reply to those comments improve others’ experiences conversing in that forum? Is it more helpful for others to publicly, but benevolently (with a polite tone, demonstrated understanding of the original comment, and empathy for the commenter; Young Reusser et al., 2021), correct the post? Is going along with or joking along with the commenter in a benevolent way helpful? Or is retaliating – returning toxicity for toxicity – the best strategy? Using real Reddit conversation pairs – a toxic comment followed by a reply – as stimuli, we conducted a pilot study (*n* = 126 participants) and propose an experiment (proposed *n* = 1122 participants) investigating the impact of three kinds of replies to online toxicity (benevolent correction, benevolent going-along, or retaliation) on observers’ self-reported freedom to contribute to the conversation, their belief that the toxicity will be reduced, and their overall impression that justice has been restored. We found evidence that… These findings suggest…

*Keywords:* online discourse, benevolence, empathy, forgiveness

# Responding to Online Toxicity: Which Strategies Make Others Feel Freer to Contribute, Believe That Toxicity Will Decrease, and Believe that Justice Has Been Restored?

You may have had the experience of reading through an engaging, lively discussion online and suddenly coming across a post that is toxic; in other words, hateful, aggressive, or disrespectful, potentially making you, and others, want to leave that discussion (Perspective, 2021). According to Pew Research, many U.S. adults have had similar experiences; 66% of U.S. respondents reported witnessing harassment online, from less severe cases (e.g., offensive name-calling) to more severe (e.g., physical threats; Pew Research Center, 2017). Forty-one percent of respondents reported personally experiencing such harassment (Pew Research Center, 2021). Pew also found that many U.S. adults prefer strict consequences for such users, with 51% of respondents saying permanently banning those who bully or harass others would be very effective in reducing the behavior. Many respondents (60%), though, say that those who witness harassing behavior should take on a “major role” in fixing the issue (Pew Research Center, 2017).

Toxic behavior online, then, is common and widely regarded as a problem that needs to be addressed, not only by platforms but by users themselves. Many researchers have studied online toxicity, identifying it (e.g., Wulczyn, Thain and Dixon, 2017; Xia, Zhu, Lu, Zhang and Gu, 2020) and highlighting factors that can lead to it (e.g., Almerekhi, Kwak, Salminen and Jansen, 2020). Another question, though, and one that has received less attention, is whether prosocial, positive responses to toxic behavior can help reduce toxicity.

Bao et al. (2021) asked participants to view hundreds of pairs of conversations from the discussion platform Reddit and select which of the two was more prosocial. They found that prosocial conversations tended to share information more, were rated higher by the community, received more engagement (e.g., total replies, sustained conversation from the same users), and included more polite content (e.g., compliments, laughter). Prosocial conversations were not simply, according to Bao et al. (2021), “the absence of antisocial behavior” (p. 1138). They found a small-to-moderate negative correlation between the likelihood that a conversation was prosocial and the number of toxic replies in that conversation, suggesting that kindness in response to toxicity might help conversations develop with a modestly kinder trajectory.

In previous work by the current research team (Young Reusser, Veit, Gassin, Case and Reusser, 2021), we sought to understand prosocial behavior in the face of toxicity online. We developed a three-item scale to measure a construct we called reply benevolence. A benevolent post demonstrates understanding of the content of the original comment, empathy (care for, interest in, respect for, and concern for the well-being of the original commenter), and is politely, thoughtfully, and/or helpfully worded (Young Reusser et al., 2021). We asked 792 online volunteers to rate the benevolence of a single reply to the roughly 8,600 most-toxic comments in a dataset of about 11 million Reddit posts from the month of January, 2016 using this scale. We found that 37.83% of the replies were rated above the scale midpoint in benevolence. Benevolence in response to highly-toxic comments was a common strategy.

Since publishing that report, we explored that same Reddit dataset further to see what sorts of replies counted as benevolent. Our reasoning was that someone could appear empathic, understanding and polite because they are simply joking along with, rather than trying to counteract, the toxic commenter. In brief, we identified two distinct strategies in the 669 most-benevolent replies to toxic posts in our data: 50.24% of these replies corrected the toxic comment in some way (Benevolent Correction) and 37.10% of the replies went along with (that is, agreed with or joked along with) the toxic commenter (Benevolent Going-Along). These strategies were strongly negatively correlated (*r* = -.74, *p* < .001; see Supplemental Materials for further details).

Building on this descriptive work, we were interested in the current research in understanding any potential differences in effectiveness among three types of replies to toxic posts online: Benevolent Corrections, Benevolently Going-Along, and another understandable response to toxicity – retaliation (Retaliatory). We assessed three separate dependent measures we thought might be impacted by these replies: 1) how free participants feel to engage in the conversation, 2) their sense that toxicity will decrease, and 3) their sense that justice has been restored.

**Dependent Measure One. How Free Do People Feel to Engage in the Conversation?**

Our first research question is as follows: To what extent do Benevolently Correcting, Benevolently Going-Along, or Retaliatory responses to toxic comments online make observers feel freer to contribute to the conversation?

The Spiral of Silence Theory (Noelle-Neumann, 1977) holds that in situations where there is disagreement about public opinion, individual speakers will tend to first identify what position they are able to express while avoiding social isolation. If they realize that an expressed opinion doesn’t have much social support, they will tend to express that opinion less and less. Perceiving that others are “with them,” though, can make speakers more likely to contribute in a public forum. For example, Zerback and Fawzi (2017) found that participants who supported evicting immigrants were substantially more likely to post a comment to a fictitious Facebook group if previous commenters agreed with their position compared to if previous commenters disagreed with them. Those who opposed evicting immigrants, though, did not change their commenting behavior depending on prior comments.

According to Spiral of Silence Theory, a toxic comment may serve to indicate to a forum user that at least some other users are actively aggressive. This may signal that posting one’s own opinion will result in further social isolation (here, possibly, in the form of an attack from the toxic commenter). If the toxic comment is left uncorrected, participants could feel less free to engage in the forum. This is consistent with some research on online toxicity; for instance, Mohan et al. (2017) found that the higher the percentage of toxic posts in a Reddit forum, the fewer posts and fewer unique users there were (see Salehabadi (2019) for a similar finding based on Twitter conversations). In addition, a 2015 Harassment Survey by the Wikimedia Support & Safety Team suggested that while half (51%) of respondents who witnessed personal attacks or harassment said it did not change their involvement in the community, 42% at least considered not contributing to the site, with only 4% saying their contributions increased (Wikimedia Support & Safety Team, 2015).

Other work suggests, though, that counter to Spiral of Silence Theory, toxic comments may increase forum engagement. Xia et al. (2020) found a small positive correlation between the toxicity of a Reddit comment and the number of direct replies to that comment across several topics. In their analyses of BBC news-related message boards, Chmiel et al. (2011) found that boards with more negative emotional content had more user engagement. They speculated, though, that angry exchanges “may encourage other users to adopt a similar tone” (p. 14). In other words, more engagement does not necessarily mean healthier conversation. Consistent with this, Xia et al. (2020) found that Reddit replies to toxic comments were slightly more likely to be toxic themselves. While Kolhatkar and Taboada (2017) found evidence that the toxicity of a comment was unrelated to that comment’s ability to promote civil dialogue, they based this on direct replies to news articles. Their findings may not be applicable to interpersonal online conversation.

How might direct replies to toxic comments impact how free users feel to contribute, according to Spiral of Silence Theory? A reply that corrects the toxicity in a benevolent way might serve as a signal that the comment was an unpopular one, increasing the likelihood that users will speak up. A retaliatory reply might serve a similar function, setting a norm for others of speaking out against the toxicity, albeit in a more negative way. A reply that is benevolent but that simply goes along with the toxicity might actually serve to reinforce a sense that others *agree* with the toxic commenter, potentially *decreasing* the likelihood that the participant will speak up.

Research from the forgiveness literature supports the idea that publicly correcting toxicity could free others up to contribute. who experienced more psychological forgiveness (forgiving, wanting good things to happen to them and for others to treat them fairly) for the uncivil coworker. Ravedwasltid.

Hypothesis 1. Participants will feel freer to contribute to a conversation initiated by a specific toxic comment after a Benevolent Correction or Retaliatory reply compared to a reply that Benevolently Goes Along with the toxic comment.

**Dependent Measure 2. Will Toxicity Decrease?**

Cialdini, Kallgren and Reno (1991) distinguished in their Focus Theory of Normative Conduct between descriptive norms, social norms which highlight what people typically do in a situation, and injunctive norms, those which highlight what people believe is appropriate. In a classic example of this (Cialdini, Reno & Kallgren, 1990), the likelihood that participants would litter was highest when a descriptive norm was highlighted; a confederate dropping litter in a littered environment suggests that littering is what people typically do. The likelihood of littering was lowest when an injunctive norm was highlighted; a confederate dropping litter in a clean environment suggests that littering is socially unacceptable.

What sorts of replies to online toxicity will make it appear socially unacceptable and thus less likely to occur? A person correcting a toxic post in a benevolent way might highlight two injunctive norms: that saying toxic things is wrong and treating others kindly is right. The injunctive norm to be kind to others might be especially salient to observers given how unexpected it might seem in that situation. If observers intuit this, they could believe the toxic commenter will be less likely to reoffend after a benevolent correction. A retaliatory reply could highlight the injunctive norm that saying toxic things is wrong, but might not highlight the importance of treating others kindly. Observers in this case might believe the toxic commenter will be less likely to reoffend, but perhaps not to the same extent. A reply that is benevolent but goes along with the toxic comment might not highlight injunctive norms at all, leading observers to feel that the toxic commenter’s behavior will not change (or possibly will increase due to an injunctive norm that toxicity is socially acceptable here). Theoretically, then, benevolent corrections should be most effective for dissuading toxicity, and benevolently going along with the toxicity least effective, with retaliation somewhere in the middle.

Some evidence that corrective posts might dissuade toxicity exists. Hangartner et al. (2021) found that politely-worded corrective messages intended to elicit empathy from Twitter users who posted hate speech resulted in a small reduction in the amount of hate speech posted and slightly increased the likelihood that that user would delete their negative tweet. In an elementary school context, Saarento et al. (2013) found that when teachers were perceived to disapprove of bullying, students self-reported less victimization by other students. On the other hand, found that when Tusers,r

What about retaliation? Work by Molnar, Chaudhry and Loewenstein (2020) is consistent with the idea that retaliation could be viewed as a deterrent for future toxicity. When their participants learned that a target individual was unfair to them, the majority of those who retaliated monetarily preferred to include a message explaining the reason (e.g., “because you were unfair to your partner” (p. 7)). Further, in another experiment, participants who retaliated and explained their reasoning expected the target individual to treat others better in the future – in other words, some participants thought retaliation (with explanation) would dissuade the negative behavior.

online

Participants could believe that retaliation might not be an effective deterrent, though. While Benesh et al. (2016) argue that continued, civil conversation might slowly reduce the extremity of hate speech, they specifically discourage hostility and aggression, which they argue can entrench the original commenter and escalate the situation. Indeed, Herschovis et al. (2018) found that when participants confronted incivility in the workplace, incivility was more likely to reoccur. In other words, the injunctive norm being highlighted by a retaliatory response might be more that aggression is acceptable than that toxicity is not, potentially encouraging toxicity as a result.

Hypothesis 2. Participants will believe to a greater extent that toxicity has been dissuaded…

1. when the replier Benevolently Corrects or Retaliates compared to Benevolently Going Along.
2. when the replier Benevolently Corrects compared to either alternative (Benevolently Going Along or Retaliating).

**Dependent Measure 3. Has Justice Been Restored?**

Wenzel and Okimoto (2008) distinguish between two psychological motivations for bringing about justice after a transgression: retributive and restorative. In retributive justice, moral order which has been disrupted by a perpetrator is reestablished by punishing them and giving them what they deserve. The punishment fixes the problem, regardless of the transgressor’s remorse or lack thereof. In restorative justice, the transgressor is viewed as in conflict with the victim (or victims) and the community as a result of their actions. The goal of restorative justice is for all parties to gain “a shared understanding of the harm the offense has done and the values it violated” (p. 378). Importantly, while transgressors should accept accountability for their actions in this model, victims are also urged to see transgressors with benevolence, forgiving them and viewing them as “a morally worthy person capable of more than wrongdoing” (Govier, 1999, p. 60).

Accordingly, which sorts of replies to an online transgression – a toxic comment – will give participants the sense that justice has been restored? Perhaps a restorative-justice approach, correcting the comment in a benevolent way, will do so. When participants were asked to imagine that another student had lied to them to get out of a group project, their sense that the situation was fair and just was moderately higher if they imagined forgiving that student than if they did not, regardless of whether the offender apologized (Wenzel and Okimoto, 2010). Forgiving also reduced hostile emotions compared to not forgiving. If forgiveness is analogous to kindly, empathically correcting toxicity, perhaps observers might see benevolent correction as increasing fairness and justice.

Alternatively, a retributive-justice approach, a retaliatory response, could be seen as more restorative of justice. Liang et al. (2018) found that when participants were asked to stab an online voodoo doll representing a rude or otherwise hostile supervisor, their implicit sense of injustice was lower than participants who did not retaliate in this way. Wang and Todd (2021) found that when participants read a brief conversation between a negative target individual (a member of a White supremacist group) and another person who either condemned the target or empathized with them, participants had more respect for this second person if they condemned the target.

Hypothesis 3. After reading a set of conversations, participants will feel that justice has been restored…

1. after a Benevolent Correction compared to either a Benevolently Going-Along or Retaliatory reply.
2. after a Retaliatory reply compared to a Benevolent reply of either kind (Correction or Going Along).

We specify which condition differences would support each hypothesis as well as what would disconfirm each hypothesis in the PCI RR study design below (Table 1).

**Table 1**

***PCI RR Study Design: Sampling and Analysis Plan, Findings That Would Support or Disconfirm Each Hypothesis***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question** | **Hypothesis** | **Sampling plan** | **Analysis Plan** | **Sensitivity rationale** | **Interpretation given different outcomes** | **Theory that could be shown wrong by the outcomes** |
| Q1: To what extent do Benevolently Correcting, Benevolently Going Along, or Retaliatory responses to toxic comments online make observers feel freer to contribute to the conversation? | 1. Participants will feel freer to contribute to a conversation initiated by a specific toxic comment after a Benevolent Correction or Retaliatory reply compared to a reply that Benevolently Goes Along with the toxic comment. | According to GPower, the sample size required to detect the smallest effect size of interest (*f* = .11) 90% of the time using an ANCOVA with a three-level factor (condition) and three covariates is 1049. We estimate based on the pilot that roughly 7% might be dropped for failing an attention check, so we added an additional 7% to our proposed sample, resulting in a target sample of 1122 participants. | Participants who fail an attention check will be dropped prior to all analyses. Multilevel regression nesting ratings within comment-reply pair (1-12) and participant, predicting perceived freedom to contribute from condition (fixed factor, between-subjects; Benevolent Correction vs. benevolent Going-Along vs. Retaliatory reply) and three covariates (comfort with offensive language, first impression of toxic commenter, and willingness to self-censor). | The smallest effect size of interest of *f* = .11 was drawn from research by Zerback and Fawzi (2017) who found that size for the difference in likelihood to post a comment when the majority is on vs. opposing your side. | Support for H1: The Benevolently Going Along condition’s mean is lower than the other two at the .05 level. The two other conditions do not differ. If the Benevolently Going Along condition’s mean is similar to either other condition, this hypothesis would be disconfirmed. This hypothesis is agnostic as to the difference between the other two conditions  Note that we do not have a prediction as to the size of the effects for these hypotheses. | If H1 is supported, it would be consistent with Spiral of Silence Theory’s (Noelle-Neumann, 1977) prediction that the apparent majority opinion leads those in the apparent minority to be less and less likely to speak.  If H1 is not supported, it would suggest that in some cases, those who think they hold a minority opinion (e.g., not liking toxicity when someone has Benevolently Gone Along with a toxic comment) still want to speak up. |
| Q2: To what extent do Benevolent Corrections, Benevolent Going Along, or Retaliatory responses to toxic comments online make observers feel that the toxicity has been dissuaded? | Two possibilities:  2. Participants will believe to a greater extent that toxicity has been dissuaded…   1. when the replier Benevolently Corrects or Retaliates compared to Benevolently Going Along. 2. when the replier Benevolently Corrects compared to either alternative (Benevolently Going Along or Retaliating). | See above | Parallel multilevel regression analysis to Q1, with perceptions that toxicity has been dissuaded as the dependent measure. |  | Support for H2a: The Benevolently Going Along condition’s mean is lower than the other two at the .05 level. This hypothesis is agnostic as to the difference between the other conditions. If the Benevolently Going Along condition is similar to one or both of the other two conditions, this hypothesis would be disconfirmed.  Support for H2b: The Benevolent Correction condition’s mean is higher than the other two at the .05 level. This hypothesis is agnostic as to the difference between the other conditions. If the Benevolent Correction condition is similar to one or both of the other two conditions, this hypothesis would be disconfirmed.  Note that we do not have a prediction as to the size of the effects for these hypotheses. | If H2a is supported, it would be consistent with Cialdini et al.’s (1991) Focus Theory of Normative Conduct, potentially indicating that any correction of toxicity highlights an injunctive norm that toxicity is not socially acceptable, dissuading toxic posts.  If H2b is supported, it would suggest that for participants to think toxicity will be dissuaded, the injunctive norm cannot only be that toxicity is not acceptable, but positively that behaving kindly is socially expected.  If neither is supported, it would be inconsistent with the Theory of Normative Conduct. |
| Q3: After reading a set of conversations, to what extent will observers feel that justice has been restored by Benevolent Corrections, Benevolent Going Along, or Retaliations? | Two possibilities:   1. Benevolent Corrections (vs. Benevolently Going Along or Retaliating) will make participants feel more that justice has been restored 2. Retaliatory responses (vs. Benevolently Correcting or Going Along) will make participants feel more that justice has been restored | See above | ANCOVA predicting perceived justice restoration from condition (between-subjects; Benevolent Correction vs. Benevolent Going Along vs. Retaliatory reply) and two covariates (comfort with offensive language and willingness to self-censor) |  | Support for H3a: The Benevolent Correction condition’s mean is higher than the other two at the .05 level. This hypothesis is agnostic as to the difference between the other conditions. If the Benevolent Correction condition’s mean is similar to either of the other two conditions, this hypothesis would be disconfirmed.  Support for H3b: The Retaliatory condition’s mean is higher than the other two at the .05 level. This hypothesis is agnostic as to the difference between the other conditions. If the Retaliatory condition is similar to either of the other conditions, this hypothesis would be disconfirmed.  Note that we do not have a prediction as to the size of the effects for these hypotheses. | If H3a is supported, it would suggest that participants are taking a restorative-justice approach to the online conflict.  If H3b is supported, it would suggest that participants are taking a retributive-justice approach. |
| Manipulation check - benevolence | Ensure that the benevolent replies are rated as more benevolent than the retaliatory replies |  | ANOVA predicting overall rated benevolence of a set of replies from condition (Benevolent Correction vs. Benevolent Going-Along vs. Retaliatory) |  | The two Benevolent conditions should be rated as more benevolent than the Retaliatory condition at the .05 level. If not, this would undermine our ability to interpret analyses related to RQ1, RQ2 and RQ3. |  |
| Manipulation check - correcting | Ensure that the Benevolently Correcting replies are rated as more correcting than the benevolently Going Along replies |  | ANOVA predicting overall rating of a set of replies’ attempt to correct the initial toxic comment from condition (Benevolent Correction vs. Benevolent Going-Along vs. Retaliatory) |  | The Benevolent Correction condition should be rated as more correcting of the initial comment than the Benevolently Going Along condition at the .05 level. If not, this would call into question any analyses which suggest a difference between the benevolent correction and benevolent going-along conditions. |  |
| Manipulation check – retaliatory | Ensure that the Retaliatory replies are rated as more retaliatory than either other condition |  | ANOVA predicting overall ratings of a set of replies’ retaliatory ratings from condition (Benevolent Correction vs. Benevolent Going-Along vs. Retaliatory) |  | The Retaliatory condition should be rated as more retaliatory than the other two at the .05 level. |  |
| Manipulation check – first impression of toxic commenter (pilot); perceived toxicity of initial comment (proposed experiment) | Ensure the participant’s first impression toxic commenter (pilot) or the perceived toxicity of the initial comment (proposed experiment) is similar across conditions. If this is not the case, first impression will be controlled for. |  | Multilevel regression nesting ratings within comment-reply pair (1-12) and participant, predicting first impression from condition (fixed factor, between-subjects; Benevolent Correction vs. Benevolent Going Along vs. Retaliatory reply). |  | If the first impression of each toxic commenter differs by condition at the .05 level, this will be included as a covariate in the proposed analyses for RQ1, RQ2 and RQ3. |  |

**Pilot Study**

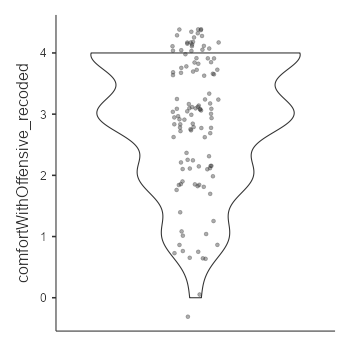
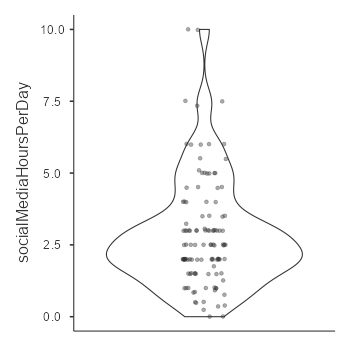
***Method***

We conducted a pilot study to estimate the smallest effect size of interest for use in a power analysis to determine an appropriate sample size for our proposed experiment. We also hoped to refine our materials and methods based on this study. The methods and materials used in the pilot are almost identical to those used in the proposed experiment, so we devote more time to them here. The pilot tested the same hypotheses as the proposed experiment. A pdf of our Qualtrics survey (pilotQualtrics.pdf) and deidentified pilot data can be found on the Open Science Framework (<https://osf.io/6dwjx/?view_only=2b45b35cf37e46e5818a40bf79fc981d>).

**Participants.** We collected pilot data from 126 participants recruited from psychology courses across two faith-based liberal arts colleges (college A, *n* = 66; college B, *n* = 60) in Fall of 2021 and Spring of 2022 to test our materials and design. Nine were dropped for failing an attention check, resulting in a final sample size of 117. The average social media usage per day in our sample was 2.87 hours (*SD* = 1.90; min = 0; max = 10) and the average comfort with offensive language on a scale from 0 (it makes me very uncomfortable) to 4 (it doesn’t bother me) was 2.89 (*SD* = 1.07; min = 0; max = 4; see Figure 1). Participants were randomly assigned to either the Benevolent Correction condition (*n* = 40 after two were dropped), the Benevolent Going Along condition (*n* = 35 after six were dropped), or the Retaliatory condition (*n* = 42 after one was dropped). Using a chi-square goodness-of-fit test, we did not find evidence that significantly more participants were dropped from a given condition, *χ*²(2) = 4.67, *p* = .097.

**Figure 1**

*Distributions of Social Media Use and Comfort with Offensive Language in Pilot Study*



Comfort with Offensive Language

Social Media Use (hours per day)

*Note*. These violin plots represent the frequency distributions of the 117 pilot participants’ social media use in hours per day and comfort with offensive language (0 = it makes me very uncomfortable; 4 = it doesn’t bother me).

**Materials.**

***Reddit Conversations.*** Participants in each condition read four separate conversations during the experiment in a randomized order. These were selected from Young Reusser et al.’s (2021) Reddit 2016 dataset. They were taken verbatim from the dataset and unedited, preserving any grammatical errors or typos. Each consisted of a toxic comment followed by a single reply – a comment-reply pair. All benevolent replies were rated as a 5 or above on the 1-6 benevolence scale described in Young Reusser et al. (2021). Three of the four benevolently correcting replies were rated above the scale midpoint in Correcting by undergraduate research assistants; one was researcher-selected. Three of the four benevolently going-along replies were rated above the scale midpoint in Going Along; one was researcher-selected. All four retaliatory replies were researcher-selected from a list of the least-benevolent replies and further, demonstrated a negative, aggressive, dismissive and/or rude tone. Comments which revealed a controversial opinion (e.g., a political argument) were excluded from consideration. We also tried to keep the lengths of the comments and replies manageable and similar across conditions, although since these were real conversations, we could not exactly match comment length. All conversations, their ratings and word counts can be found in Appendix A.

***Per-Pair Ratings.*** After each comment-reply pair, participants rated their first impression of the toxic commenter using seven options from -3 (Very negative) to +3 (Very positive). They were then asked to consider both individuals in the conversation and indicate how free they felt to contribute using three items (e.g., “How likely would you be to contribute to this conversation?”) written by the experimenters. We dropped one reverse-worded item (“If you were to post to this forum, to what extent would you feel the need to hide what you really think from the rest of the group?”) to improve the internal consistency of the scale from .65 to .70 and averaged responses to the remaining two items together. Participants responded using seven options from 0 (Not at all) to 6 (Very likely).

Participants then reported the extent to which they believed the reply addressed their concerns and discouraged the toxicity of the initial comment using four items (e.g., “The response is an appropriate way to address the toxicity of the first comment;” “The response will discourage the first commenter from continuing to post in the same negative tone as before”, Cronbach’s *α* = .88). They used seven options from -3 (Strongly disagree) to +3 (Strongly agree). These items can be found in Appendix B and means and standard deviations for these measures can be found in Table 2.

| **Table 2**  ***Means and Standard Deviations for Key Variables in Pilot and Proposed Experiment*** | | | |  |
| --- | --- | --- | --- | --- |
|  | Pilot study  (*n* = 117) | Proposed experiment  (*n* = #) | |
| First impression of toxic commenter | -1.61 (1.11) | # (#) |  |  |
| Free to contribute per comment-reply pair | 1.92 (1.48) | # (#) |  |  |
| Toxicity addressed/dissuaded per comment-reply pair | -1.20 (1.50) | # (#) |  |  |
| Justice restored | -1.19 (1.28) | # (#) |  |  |
| Willingness to self-censor | 3.11 (0.78) | # (#) |  |  |
| Comfort with offensive language | 2.89 (1.07) | # (#) |  |  |
| Social media hours per day | 2.87 (1.90) | # (#) |  |  |

***Overall Ratings.*** After reading all four conversations, participants completed an attention check where they were asked to select from four options something they had read in one of the previous conversations. They then rated their overall impression of whether the replies to the toxic comments they had read had restored justice using seven items adapted from Wenzel, Okimoto, Feather and Platow, 2010; e.g., “The resolution to the situation is fair,” Cronbach’s *α* = .91). All items except one were identical to the original scale, with one exception: the item “The resolution has restored justice” was modified to “The replies have restored justice” to make it more specific to this study. Participants responded using seven options from -3 (strongly disagree) to +3 (strongly agree).

***Individual Differences.*** Participants then reported their Willingness to Self-Censor (Cronbach’s *α* = .83), an eight-item scale developed by Hayes, Glynn and Shanahan (2005; e.g., “It is difficult for me to express my opinion if I think others won’t agree with what I say”) using five options from 1 (strongly disagree) to 5 (strongly agree). They rated their level of comfort with reading offensive language from 0 (it makes me very uncomfortable) to 4 (it doesn’t bother me at all) and then estimated the total time in hours and minutes in an average day they were on social media for personal use. These were included so they could be controlled for in our analyses; a participant’s tendency to avoid stating their true opinion could relate to how free they feel to contribute to any conversation and whether they feel that replying at all is a helpful strategy. Their level of comfort with offensive language could relate to their perception of the need to correct a toxic comment. Social media use was included to help characterize our sample (e.g., are these people who are used to online conversation?).

**Procedure.** After random assignment to condition, participants read through one toxic comment and one reply from four separate Reddit conversations in a randomized order. They provided ratings of each conversation (first impression of the toxic commenter, free to contribute (3 items, order randomized) and toxicity addressed (4 items, order randomized)), then an attention check, then their overall impression of the fairness/justice of the resolution (7 items, order not randomized). They then completed individual difference measures (Willingness to Self-Censor (8 items, order not randomized), comfort with offensive language and total time per day on social media), and were debriefed. 90% of participants finished in 30 minutes or less. This study had IRB approval from Olivet Nazarene University and Houghton University.

***Results***

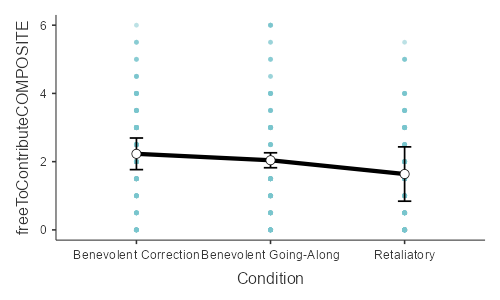
**Per-Pair Ratings.** Our first analyses were on the ratings of each separate conversation. These analyses were all multilevel regression models nesting ratings within pair (1-12; 4 each across three conditions) and participant. Condition (benevolent correction s. benevolent going-along vs. retaliatory) was a between-subjects fixed factor in each model. Each analysis involved four ratings per 117 people, or 468 observations. We maintained the false discovery rate at 5% across all analyses using the Benjamini-Hochberg false discovery procedure (Benjamini and Hochberg ,1995).

***First Impression of Toxic Commenter.*** To ensure that the toxicity of the initial comments was consistent across all three conditions, we conducted a multilevel regression predicting the first impression of the toxic commenter from condition (Intraclass Correlation Coefficient (ICC) = 0.08. This specifies the proportion of variation in first impressions explained by the cluster variables, conversation pair (1-12) and participant). The effect of condition was not significant, *F*(2, 114) = 2.53, *p* = .145; nor was the difference between Benevolent Going Along (*M* = -1.58, *SE* = 0.11, 95% *CI* [-1.79, -1.37]) and Benevolent Correction (*M* = -1.37, *SE* = 0.10, 95% *CI* [-1.57, -1.17]) conditions (planned comparison *t*(114) = -1.44, *p* = .152). However, given that the confidence intervals for the direct comparison of the Retaliatory (*M* = -1.63, *SE* = 0.10, 95% *CI* [-1.82, -1.44]) and Benevolent Correction (planned comparison *t*(114) = -1.89, *p* = .061) only overlapped by .07 first impression scale points,, we decided to control for the first impression in all multilevel analyses in the pilot study. Note that the following analyses were conducted both including the covariates (first impression of the toxic commenter, willingness to self-censor, and comfort with offensive language) and without, and the effect of condition is reported for both.

***Free to contribute.*** Using a multilevel model predicting how free participants felt to contribute to the conversation controlling for the first impression of the toxic commenter, willingness to self-censor, and comfort with offensive language (ICC = 0.02), we did not find a significant difference among conditions, *F*(2, 8.21) = 2.17, *p* = .18; without covariates, *p* = .14; see Figure 2. Planned comparisons suggested that the difference between the Benevolent Correction (*M* = 2.23, *SE* = 0.15, 95% *CI* [1.77, 2.69]) and the Retaliatory condition (*M* = 1.64, *SE* = 0.25, 95% *CI* [0.84, 2.43]) was large, albeit non-significant, *t*(4.76) = -2.07, *p* = .096, *r* = .69; without covariates, *p* = .065. The comparison between the Benevolent Going Along and Benevolent Correction conditions was moderate but non-significant, *t*(7.58) = -1.03, *p* = .34, *r* = .35; without covariates, *p* = .24. A post hoc comparison between the Benevolent Going Along and Retaliatory conditions was large but non-significant, *t*(5.63) = 1.02, *pBonferroni* = 1.00, *r* = .59; without covariates, *pBonferroni* = .55. Willingness to self-censor[[1]](#endnote-1) was negatively related to how free participants felt to contribute, *b* = -0.62, *SE* = 0.08, 95% *CI* [-0.79, -0.46], *t*(447.91) = -7.40, *p* < .001. Neither comfort with offensive language (*b* = 0.08, *SE* = 0.06, 95% *CI* [-0.04. 0.20], *t*(447.99) = 1.25, *p* = .21) nor first impression of the toxic commenter (*b* = 0.05, *SE* = 0.06, 95% *CI* [-0.06. 0.17], *t*(435.39) = 0.94, *p* = .35) were significant.[[2]](#endnote-2)

**Figure 2**

*Free to Contribute Across Conditions in Pilot*



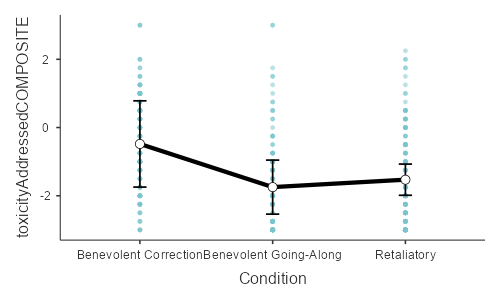
Free to Contribute

*Note.* Error bars represent 95% confidence intervals. Circles represent condition means.

***Toxicity addressed/dissuaded.*** Using a multilevel model predicting the extent to which participants felt the reply addressed and discouraged the initial toxicity controlling for the first impression of the toxic commenter, willingness to self-censor, and comfort with offensive language (ICC = 0.28), we again failed to find a significant difference among conditions, *F*(2, 4.42) = 3.79, *p* = .11; without covariates, *p* = .070; see Figure 3.[[3]](#endnote-3) Planned comparisons suggested that the difference between the benevolent correction (*M* = -0.48, *SE* = 0.40, 95% *CI* [-1.75, 0.78]) and benevolent going along (*M* = -1.75, *SE* = 0.25, 95% *CI* [-2.54, -0.95]) conditions was large, *t*(5.05) = -2.70, *p* = .042, r = .77; without covariates, *p* = .026, but not significant according to the Benjamini-Hochberg FDR procedure. The difference between the benevolent correction and retaliatory (*M* = -1.53, *SE* = 0.14, 95% *CI* [-1.99, -1.07]) conditions was also large, *t*(3.78) = -2.48, *p* = .072, *r* = .79, though non-significant (without covariates, *p* = .045). A post hoc comparison suggested that the difference between the benevolent going-along and retaliatory conditions was moderate but not significant, *t*(4.77) = -0.76, *p*Bonferonni= 1.00, *r* = .33. Again, willingness to self-censor was negatively related to toxicity addressed, *b* = -0.19, *SE* = 0.08, 95% *CI* [-0.34, -0.03], *t*(444.11) = -2.40, *p* = .017, though this did not reach significance after the Benjamini-Hochberg FDR procedure. First impression of the toxic commenter was positively related to toxicity addressed, *b* = 0.16, *SE* = 0.05, 95% *CI* [0.05, 0.26], *t*(327.12) = 2.92, *p* = .004; the less negative (more positive) the toxic comment was perceived to be, the more participants thought the reply addressed the toxicity. Comfort with offensive language was not related to toxicity addressed, *b* = 0.07, *SE* = 0.06, 95% *CI* [-0.04, 0.18], *t*(444.21) = 1.21, *p* = .23.

**Figure 3**

*Toxicity Addressed/Dissuaded Across Conditions in Pilot*



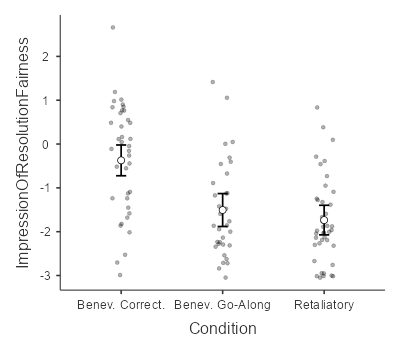
Toxicity addressed/dissuaded

*Note.* Error bars represent 95% confidence intervals. Scale reflects opinion that toxicity has been addressed/dissuaded from -3 (Strongly disagree) to +3 (Strongly agree).

**Overall Ratings: Justice Restored**. Recall that after reading all four conversations, each participant provided their overall impression of whether the replies resolved the situation fairly. We conducted a between-subjects ANCOVA predicting this fair/just resolution rating from condition controlling for willingness to self-censor and comfort with offensive language. We found evidence of at least one difference among the condition means, *F*(2, 107) = 17.09, *p* < .001, η2 = 0.23, without covariates, *p* < .001; see Figure 4. Planned comparisons suggested, consistent with Hypothesis 3a, that benevolent corrections were rated as providing a more just/fair resolution (*M* = -0.37, *SE* = 0.18, 95% *CI* [-0.72, -0.02]) than benevolent going-along (*M* = -1.51, *SE* = 0.19, 95% *CI* [-1.88, -1.13]), *t*(107) = -4.37, *p* < .001, *d* = 1.04 (without covariates, *p* < .001); and retaliatory (*M* = -1.74, *SE* = 0.17, 95% *CI* [-2.07, -1.40]) replies, *t*(107) = -5.53, *p* < .001, *d* = 1.25 (without covariates, *p* < .001). We did not find evidence that the benevolent going-along and retaliatory replies differed (*pTukey* = .64; without covariates, *pTukey* = .61). Comfort with offensive language was not related to fair/just resolution ratings, *F*(1, 107) = 3.72, *p* = .057, η2 = 0.03 and willingness to self-censor was not a significant predictor, *F*(1, 107) = 1.47, *p* = .23, η2 = 0.03.

**Figure 4**

Overall Justice Restored Across Conditions in Pilot



Justice Restored

*Note.* Error bars represent 95% confidence intervals. Scale reflects opinion that justice has been restored from -3 (Strongly disagree) to +3 (Strongly agree).

***Discussion***

Of our three primary dependent measures, we only found evidence of a condition effect for participants’ overall ratings of the extent to which the replies to toxic comments restored justice. Consistent with Hypothesis 3a, when a Redditor replied in a benevolent way and corrected the initial toxic remark, participants felt justice had been restored more than either of the other two kinds of replies, which did not differ. The size of this effect was large. We did not find support in this pilot for either Hypothesis 1, regarding how free participants felt to contribute to the conversation, or Hypotheses 2a and 2b, related to perceptions that toxicity had been dissuaded.

***Smallest Effect Size of Interest***

To estimate the smallest effect size of interest for our proposed experiment, we used effect sizes from our pilot study as well as some effect sizes from the literature. The pilot effect sizes of interest related to condition differences for the three key dependent variables: how free participants felt to contribute (r = .35 to r = .69), the extent to which they thought toxicity would be dissuaded (r = .33 to r = .79), and their overall sense that justice had been restored (d = 1.04 to d = 1.25). Willingness to speak out in online settings, in Porten-Chee and Eilders (2015), was higher when personal opinion and the opinion climate clashed (*β* = .14 and *β* = .23), though they used a different measure than we plan to. Zerback and Fawzi (2017) similarly found an effect size of *r* = .11 for the difference in likelihood to post a comment when the group majority is on your side vs. opposing your position. One example in the literature on reducing toxicity suggests that actual posts become less toxic after humor is used (d = .38 and d = .36), though this isn’t a measure of perceived toxicity reduction (Elsayed and Hollingshead, 2022). Strelan, Di Fiore, and Van Prooijen (2016) found an effect size of *d* = .93 for the difference in perceived justice between conditions where participants punish vs. cannot punish a transgressor. Schoenebeck, Kaimson and Nakamura (2021) found an effect size of *g* = 1.07 for the perceived justice of mediation versus banning as a response to online toxicity targeting the participant. Taken together, we decided to use *r* = .11 (*f* = .11) as our smallest effect size of interest.

A power analysis using G\*Power for an ANCOVA with three conditions, three covariates, an effect size of *f* = .11, a family-wise alpha of .05, and 90% power suggested a total sample size of 1049. Nine of the 126 pilot participants were dropped for failing an attention check, or 7% of the sample. We added an additional 7% to our proposed sample size, resulting in a target sample of 1122 participants. We will use the Benjamini-Hochberg procedure to keep the false discovery rate at 5% (Benjamini and Hochberg, 1995).

***Method***

**Participants.** We plan to recruit 800 participants using CloudResearch’s Mechanical Turk Toolkit (Litman, Robinson and Abberbock, 2017) to complete our experiment. We will retain any beyond that number who participate. Participants who do not complete any of the key measures will be dropped prior to analysis. Those with partial data will be retained, though if they are missing data for a variable in an analysis, their data will not be included in that particular analysis. Those who answer a subset of questions for a multi-item scale will be given the average of the items they completed as their composite score. Participants will be paid $1.00 to participate and we anticipate based on the pilot that the study will take roughly 15-30 minutes. We will drop any participants who fail an attention check. **[averages and standard deviations for social media usage per day and comfort with offensive language will be reported here. Violin plots of these variables will be presented in Figure 5].** Participants were randomly assigned to either the Benevolent Correction condition (*n* = # after \_\_\_ were dropped), the Benevolent Going Along condition (*n* = # after \_\_\_ were dropped), or the Retaliatory condition (*n* = # after \_\_\_\_ was dropped). **[we will report here the results of a chi-square goodness-of-fit test comparing the drop rates across condition].** Our proposed experiment has already been programmed in Qualtrics and recruitment has been set up on the CloudResearch platform, so data collection can begin immediately upon approval.

**Figure 5**

*Distributions of Social Media Use and Comfort With Offensive Language in Proposed Experiment*

**[Figure 5 will go here]**

*Note*. These violin plots represent the # participants’ social media use in hours per day and comfort with offensive language (0 = it makes me very uncomfortable; 4 = it doesn’t bother me).

**Modifications to Pilot Procedure and Materials.** Our proposed experiment’s methods draw heavily from our pilot study, with several modifications. First, we will replace one of the three items intended to measure participants’ freedom to contribute (“If you were to post in this forum, to what extent would you feel the need to hide what you really think from the rest of the group?”) because it did not correlate strongly with the other two in the pilot. The new question is adapted from Hampton, Shin and Lu (2017). The original item wording was “If the topic of the government’s surveillance program came up, would you be very willing, somewhat willing, somewhat unwilling, or very unwilling to join the conversation?” To match the response options given for the other two items, we modified the wording to “How willing would you be to join this conversation?”

Second, we will replace two of the twelve Reddit comment-reply pairs used in the pilot (see Appendix A). Pair 2 in the benevolent correction condition was coded relatively low by undergraduate research assistants in the extent to which the reply corrected the initial comment, so we plan to replace it with a pair in which the reply was rated as more correcting. Pair 9 will be replaced because pilot participants’ first impression of the toxic commenter was relatively positive compared to the other conversations (see Appendix A).

Third, based on reviewer feedback, we will replace the first-impression question intended to verify the toxicity of the initial comment in each pair with the more specific item wording used by Google’s Perspective API when classifying the toxicity of online text (Perspective, 2021; see Appendix B). This will be named perceived toxicity.

Fourth, to make our measure of perceptions that a reply has dissuaded the toxic commenter from continuing to post in the same negative tone more specific, we will replace items one and two, which were too general (e.g., asking whether toxicity had been addressed without defining what “addressed” means), with more specific items (e.g., “The response will encourage the first commenter to post more positively in the future”). Fifth, due to experimenter error, we did not randomize the order of either the overall ratings of fair/just resolution items or the Willingness to Self-Censor scale (Hayes et al., 2005) in the pilot. We will randomize the order of the items within each.

Finally, we will add three manipulation check items after the overall ratings of fair/just resolution but before the Willingness to Self-Censor scale (Hayes et al., 2005). The first will measure the extent to which the replies to the toxic comments demonstrated benevolence for the initial commenter. The second will measure the extent to which the replies appeared to correct the initial comment, and the third will measure the extent to which the replies retaliated against the original commenter (see Appendix B).

Otherwise, the proposed experiment will use identical materials and procedure to the pilot study. In brief, participants will be randomly assigned to read four conversations (a toxic comment followed by a reply) in a randomized order from either the benevolent correct, benevolent going-along, or retaliatory condition. After each comment-reply pair, they will report their first impression of the toxic commenter, how free they feel to contribute to the conversation (3 items, order randomized) and how much they think the toxicity has been dissuaded (4 items, order randomized). After responding to all four pairs, they will complete an attention check question asking them which statement they saw in the previous conversations. They will then report their overall impression of whether justice has been restored (7 items, randomized order), complete two manipulation check questions (how correcting and how benevolent the replies in general were, order randomized), their Willingness to Self-Censor (8 items, randomized order), comfort with offensive language (0 = it makes me very uncomfortable to 4 = it doesn’t bother me at all), total time per day spent on social media for personal use, their age, gender, and race/ethnicity (open-response). Means and standard deviations for key variables will be reported in Table 2.

***Analysis Plan***

**Scale Reliability and Composites*.*** As with the pilot study, we will assess the reliability of the multiple-item measures (free to contribute, toxicity dissuaded, overall justice restored, willingness to self-censor) using Cronbach’s alpha and average the items together as long as alpha is above .7. If alpha is lower, we will drop items until it is above .7.

**Manipulation Checks.**We will first test whether the comments in the two benevolence conditions are rated as more benevolent than the retaliatory condition using a between-subjects ANOVA predicting benevolence score from condition. **[Our report will look as follows:** We found/failed to find a significant effect of condition, *F*(#, #) = #.##, *p* = .##, *η*2 = .##. Assuming significant effect, this sentence will report any significant differences across conditions.**]** Second, to test whether the benevolent correction condition conversations are perceived as more correcting that the other two, we will conduct an ANOVA predicting correcting ratings from condition. **[Our report will look as follows:** We found/failed to find a significant effect of condition, *F*(#, #) = #.##, *p* = .##, *η*2 = .##. Assuming a significant effect, this sentence will report any significant differences across conditions. This suggests that the manipulation of how benevolent and how correcting the Reddit conversations were was/was not successful. **]**

**Per-Pair Ratings.** As with the pilot, we will conduct multilevel regression models nesting ratings within pair (1-12) and participant predicting ratings of each separate conversation. Condition (benevolent correction vs. benevolent going-along vs. retaliatory) will be a between-subjects fixed factor in each model. Each analysis will involve four scores per [sample size] people, or # observations. We will use the Benjamini-Hochberg procedure to keep the false discovery rate at 5% (Benjamini and Hochberg, 1995). Analyses will be conducted both including the covariates (perceived toxicity of the initial comment (if it differs by condition at the .05 level), willingness to self-censor, and comfort with offensive language) and without, and the effect of condition will be reported for both.

***First Impression of Toxic Commenter.*** We will conduct a multilevel regression predicting the perceived toxicity of the initial comment from condition (ICC = #.##). **[Our report will look as follows:** The effect of condition was/was not significant, *F*(#, ###) = #.##, *p* = .##; sentence regarding significant condition comparisons if there are any + statement about controlling for the first impression if they are significant.**]**

***Free to Contribute.*** We will conduct a multilevel regression predicting how free participants felt to contribute to the conversation controlling for [if necessary: the perceived toxicity of the initial comment, ]willingness to self-censor[,] and comfort with offensive language (ICC = #.##). **[Our report will look as follows:** We did/did not find a significant difference among conditions, *F*(#, #.##) = #.##, *p* = .##, without covariates, p = .##; see Figure 6. Planned comparisons suggested that[, consistent with Hypothesis #,] the difference between the benevolent correction (*M* = #.##, *SE* = #.##, 95% *CI* [#.##, #.##]) and the retaliatory condition (*M* = #.##, *SE* = #.##, 95% *CI* [#.##, #.##]) was \_\_\_\_\_\_\_\_, *t*(#) = #.##, *p* = .##, *r* = .##, without covariates, *p* = .##. The comparison between the benevolent going-along and benevolent correction conditions was \_\_\_\_\_\_\_\_\_, *t*(#) = #.##, *p* = .##, *r* = .##, without covariates, *p* = .##. A post hoc comparison between the benevolent going-along and retaliatory conditions was \_\_\_\_\_\_\_\_\_\_\_, *t*(#) = #.##, *p* = .##, *r* = .##, without covariates, *p* = .##. Willingness to Self-Censor was \_\_\_\_\_\_\_\_\_\_ related to how free participants felt to contribute, *b* = #.##, *SE* = #.##, 95% *CI* [#.##, #.##], *t*(#) = #.##, *p* = .##. [perceived toxicity effect, comfort with offensive language will also be reported if significant]**]**

**Figure 6**

*Free to Contribute Across Conditions in Proposed Experiment*

**[Figure 6]**

*Note.* Error bars represent 95% confidence intervals. Circles represent condition means.

***Toxicity Dissuaded.*** We will conduct a multilevel regression predicting the extent to which participants felt the reply discouraged future toxicity controlling for [if necessary: the perceived toxicity of the initial comment, ]willingness to self-censor[,] and comfort with offensive language (ICC = #.##). **[Our report will look as follows:** We found/failed to find to find a significant difference among conditions, *F*(#, #.##) = #.##, *p* = .##, without covariates, *p* = .##; see Figure 7. Planned comparisons suggested that[, consistent with Hypothesis #,] the difference between the benevolent correction (*M* = #.##, *SE* = #.##, 95% *CI* [#.##, #.##]) and benevolent going along (*M* = #.##, *SE* = #.##, 95% *CI* [#.##, #.##]) conditions was \_\_\_\_\_\_\_, *t*(#) = #.##, *p* = .##, r = .##, without covariates, *p* = .##. The difference between the benevolent correction and retaliatory (*M* = #.##, *SE* = #.##, 95% *CI* [#.##, #.##]) conditions was \_\_\_\_\_\_\_\_\_, *t*(#) = #.##, *p* = .##, *r* = .##, without covariates, *p* = .##. A post hoc comparison suggested that the difference between the benevolent going-along and retaliatory conditions was \_\_\_\_\_\_\_\_\_\_\_, *t*(#) = #.##, *p*Bonferonni= #, *r* = .##, without covariates, *p* = .##. Willingness to Self-Censor was \_\_\_\_\_\_\_ related to toxicity dissuaded, *b* = #.##, *SE* = #.##, 95% *CI* [#.##, #.##], *t*(#) = #.##, *p* = .##. [perceived toxicity effect, comfort with offensive language will also be reported if significant]**].**

**Figure 7**

*Toxicity Dissuaded Across Conditions in Proposed Experiment*

**[Figure 7]**

*Note.* Error bars represent 95% confidence intervals. Scale reflects opinion that toxicity has been dissuaded from -3 (Strongly disagree) to +3 (Strongly agree).

**Overall ratings: Justice restored**. We will conduct a between-subjects ANCOVA predicting the overall rating of whether justice has been restored for all four conversations from condition controlling for willingness to self-censor and comfort with offensive language. **[Our report will look as follows:** We found/failed to find evidence of at least one difference among the condition means, *F*(#, ###) = #.##, *p* = .##, *η*2 = 0.##, without covariates, *p* = .##; see Figure 8. Planned comparisons suggested[, consistent with Hypothesis #,] that benevolent corrections were/were not rated as providing \_\_\_\_\_\_\_\_ just resolution (*M* = #.##, *SE* = #.##, 95% *CI* [#.##, #.##]) than benevolent going-along (*M* = #.##, *SE* = #.##, 95% *CI* [#.##, #.##]), *t*(###) = #.##, *p* = .##, *d* = #.## (without covariates, *p* = .##). They were/were not rated as providing \_\_\_\_\_\_\_\_\_ just resolution than retaliatory (*M* = #.##, *SE* = #.##, 95% *CI* [#.##, #.##]) replies, *t*(###) = #.##, *p* = .##, *d* = #.## (without covariates, *p* = .##). We did/did not find evidence that the benevolent going-along and retaliatory replies differed (*pTukey* = .##; without covariates, *pTukey* = .##). [interpretive sentence if the post hoc comparison is significant]. Comfort with offensive language was \_\_\_\_\_\_ related to just resolution ratings, *F*(#, ###) = #.##, *p* = .##, *η*2 = 0.## and willingness to self-censor was not a significant predictor (*p* = .##). **]**

**Figure 8**

*Overall Justice Restored Across Conditions in Proposed Experiment*

**[Figure 8]**

*Note.* Error bars represent 95% confidence intervals. Scale reflects opinion that justice has been restored from -3 (Strongly disagree) to +3 (Strongly agree).

***Proposed Experiment Discussion***

We will restate the findings across all analyses in plain terms and highlight which hypotheses appear to be supported or disconfirmed by our data.

**General Discussion**

We will here discuss any similarities or dissimilarities across the pilot and proposed experiment. We will also discuss the theoretical and practical implications of our data.

***Limitations and Future Directions***

We will here discuss what we know about our sample and how that limits our conclusions, talk about the limited Reddit conversation examples and the fact that conclusions might differ with different examples, consider other confounding variables we did not account for, etc. We will propose future studies to address these limitations. We will address limitations to ecological validity due to the fact that we removed conversation context from each comment-reply pair.

We might also include research suggesting that benevolently correcting toxicity might not be a very common strategy. If it ends up being relatively effective in our data, we will encourage readers to use it more. Mathew et al. (2019) found that using an empathic, kind, polite or civil tone to counter hate speech (known as counterspeech) accounted for only 9% of the YouTube comments in their sample, whereas hostile responses to hate speech occurred more frequently, about 30% of the time.

If Benevolently Going Along ends up increasing how free people feel to contribute relative to one or both of the other conditions, the following research might help explain that: Perhaps one explanation is that Benevolently Going Along, due to its positive tone, might make others feel freer to engage than a Retaliatory corrective post. Ziegele et al. (2014) found that participants “overwhelmingly preferred comments that were written in a positive and inviting tone” over negative comments. Further, Bao et al. (2021) found that conversations rated as prosocial (e.g., those that include compliments, laughter, and non-toxic content), tended to be longer than less-prosocial ones, suggesting users felt freer to contribute.

# References

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2020, April). Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of The Web Conference 2020* (pp. 3033-3040). <https://doi.org/10.1145/3366423.3380074>

Bao, J., Wu, J., Zhang, Y., Chandrasekharan, E., & Jurgens, D. (2021, April). Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021* (pp. 1134-1145), <https://doi.org/10.1145/3442381.3450122>

Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., & Wright, L. (2016). Considerations for successful counterspeech. *Dangerous Speech Project*. <https://dangerousspeech.org/wp-content/uploads/2016/10/Considerations-for-Successful-Counterspeech.pdf>

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289-300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

Chmiel, A., Sobkowicz, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M., & Hołyst, J. A. (2011). Negative emotions boost user activity at BBC forum. *Physica A: statistical mechanics and its applications*, *390*(16), 2936-2944. <https://doi.org/10.1016/j.physa.2011.03.040>

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology* (Vol. 24, pp. 201-234). Academic Press.

Cote, A. C. (2017). “I can defend myself” women’s strategies for coping with harassment while gaming online. *Games and Culture*, *12*(2), 136-155. <https://doi.org/10.1177%2F1555412015587603>

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., ... & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), https://doi.org/10.1073/pnas.2116310118 .

Hershcovis, M. S., Cameron, A. F., Gervais, L., & Bozeman, J. (2018). The effects of confrontation and avoidance coping in response to workplace incivility. *Journal of occupational health psychology*, *23*(2), 163. <http://doi.org/10.1037/ocp0000078>

Kolhatkar, V., & Taboada, M. (2017, August). Constructive language in news comments. In *Proceedings of the first workshop on abusive language online* (pp. 11-17). <http://dx.doi.org/10.18653/v1/W17-3002>

Liang, L. H., Brown, D. J., Lian, H., Hanig, S., Ferris, D. L., & Keeping, L. M. (2018). Righting a wrong: Retaliation on a voodoo doll symbolizing an abusive supervisor restores justice. *The Leadership Quarterly*, *29*(4), 443-456. <https://doi.org/10.1016/j.leaqua.2018.01.004>

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behavior Research Methods, 49(2), 433-442. <https://doi.org/10.3758/s13428-016-0727-z>

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., ... & Mukherjee, A. (2019, July). Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media* (Vol. 13, pp. 369-380). <https://ojs.aaai.org/index.php/ICWSM/article/view/3237>

Mohan, S., Guha, A., Harris, M., Popowich, F., Schuster, A., & Priebe, C. (2017, May). The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence* (pp. 51-56). Springer, Cham. <https://doi.org/10.1007/978-3-319-57351-9_6>

Molnar, A., Chaudhry, S. & Loewenstein, G. F. (2021). 'It's Not About the Money. It's About Sending a Message!': Unpacking the Components of Revenge. Available at SSRN: [http://dx.doi.org/10.2139/ssrn.3524910](https://dx.doi.org/10.2139/ssrn.3524910)

Noelle-Neumann, E. (1977). Turbulences in the climate of opinion: Methodological applications of the Spiral of Silence Theory. *Public Opinion Quarterly*, *41*(2), 143–158. <https://doi.org/10.1086/268371>

Govier, T. (1999). Forgiveness and the Unforgivable. *American Philosophical Quarterly*, *36*(1), 59-75. https://www.jstor.org/stable/20009953

Gromet, D., & Okimoto, T. (2014). Back into the Fold: The Influence of Offender Amends and Victim Forgiveness on Peer Reintegration. *Business Ethics Quarterly,* *24*(3), 411-441. <https://doi.org/10.5840/beq20147814>

Hampton, K. N., Shin, I., & Lu, W. (2017). Social media and political discussion: when online presence silences offline conversation. *Information, Communication & Society*, *20*(7), 1090-1107.

Hayes, A. F., Glynn, C. J., & Shanahan, J. (2005). Willingness to self-censor: A construct and measurement tool for public opinion research. *International Journal of Public Opinion Research*, *17*(3), 298-323. <https://doi.org/10.1093/ijpor/edh073>

Perspective (2021). Attributes and Languages. <https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages>

Pew Research Center (2017). Online Harassment 2017. <https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/>

(2021). The State of Online Harassment.

Saarento, S., Kärnä, A., Hodges, E. V., & Salmivalli, C. (2013). Student-, classroom-, and school-level risk factors for victimization. *Journal of school psychology*, *51*(3), 421-434. <https://doi.org/10.1016/j.jsp.2013.02.002>

Salehabadi, N. (2019). *The impact of toxic replies on Twitter conversations* (Doctoral dissertation, The University of Texas at Arlington).

Wang, Y. A., & Todd, A. R. (2021). Evaluations of empathizers depend on the target of empathy. *Journal of personality and social psychology*, *121*(5), 1005. <http://dx.doi.org/10.1037/pspi0000341>

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and restorative justice. *Law and human behavior*, *32*(5), 375-389. https://doi.org/10.1007/s10979-007-9116-6

Wenzel, M., & Okimoto, T. G. (2010). How acts of forgiveness restore a sense of justice: Addressing status/power and value concerns raised by transgressions. *European Journal of Social Psychology*, *40*(3), 401-417. <https://doi.org/10.1002/ejsp.629>

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2010). Justice through consensus: Shared identity and the preference for a restorative notion of justice. *European Journal of Social Psychology*, *40*(6), 909-930. <https://doi.org/10.1002/ejsp.657>

Wikimedia Support & Safety Team. 2015. Harassment Survey 2015. <https://upload.wikimedia.org/wikipedia/commons/5/52/Harassment_Survey_2015_-_Results_Report.pdf>

Wright, L., Ruths, D., Dillon, K. P., Saleem, H. M., & Benesch, S. (2017, August). Vectors for counterspeech on twitter. In *Proceedings of the first workshop on abusive language online* (pp. 57-62). <http://dx.doi.org/10.18653/v1/W17-3009>

Wulczyn, E., Thain, N., & Dixon, L. (2017, April). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web* (pp. 1391-1399). https://doi.org/10.1145/3038912.3052591

Xia, Y., Zhu, H., Lu, T., Zhang, P., & Gu, N. (2020). Exploring antecedents and consequences of toxicity in online discussions: A case study on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1-23. <https://doi.org/10.1145/3415179>

Young Reusser, A. I., Veit, K. M., Gassin, E. A., Case, J. P. and Reusser, G. M. (2021). Assessing the prevalence of benevolence in response to online toxicity on Reddit. Technology, Mind & Society 2021 Conference Proceedings, APA, <https://doi.org/10.1037/tms0000023>

Zerback, T., & Fawzi, N. (2017). Can online exemplars trigger a spiral of silence? Examining the effects of exemplar opinions on perceptions of public opinion and speaking out. *New Media & Society*, *19*(7), 1034-1051. <https://doi.org/10.1177%2F1461444815625942>

Ziegele, M., Breiner, T., & Quiring, O. (2014). What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items. *Journal of Communication*, *64*(6), 1111-1138. <https://doi.org/10.1111/jcom.12123>

**Contributions**

Contributed to conception and design: AYR, KMV, EAG, JPC

Contributed to acquisition of data: AYR, KMV

Contributed to analysis and interpretation of data: AYR

Drafted and/or revised the article: AYR

Approved the submitted version for publication: AYR, KMV, EAG, JPC

**Funding**

This work is funded by a Networking grant from the Council of Christian Colleges and Universities.

**Conflicts of Interest**

The authors have no conflicts of interest to disclose.

**Data accessibility statement**

All the stimuli, measures, participant data (for the pilot; for the proposed experiment once completed), jamovi output and analysis scripts can be found on this paper’s project page on the Open Science Framework: <https://osf.io/6dwjx/?view_only=2b45b35cf37e46e5818a40bf79fc981d>

**Appendix A: Reddit conversations used in Pilot Study and Proposed Experiment**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Condition | Pair # | Comment | Reply | Word Count | Correcting rating (1 = not at all correcting; 5 = very much) of reply | Going Along rating (1 = not at all; 5 = very much) of reply |
| Benevolent Correction | 1 | The only relation to the thread is you and your stupidity. | I understand you don’t like my opinions. But really, to keep questioning my character, etc. We just don’t agree. That’s all. | 35 | 4 |  |
| 2 | That’s how you assert dominance. If you’re an asshole. | lol, I was joking. | 13 | 3.67 |  |
| 2\* | Stop being dumb, it doesn’t matter where you use it. | Hey, insulting people isn’t necessary or valuble to the conversation. If you don’t have anything nice to say, don’t say it. | 31 | 4.67 |  |
| 3 | Your edits are dumb. | I found the edits kinda funny actually.. | 47 | 4 |  |
| 4 | That’s why I want to play a cracked version, and not pay $60 for it, idiot. | That’s fine. No need to resort to childish name calling…relax. And if you pay full price for games these days that’s just silly. You can buy this game for half that. | 11 | Researcher- selected |  |
| Benevolent Going Along | 5 | I literally cannot believe you are this stupid. | Im not one for name calling but for once I’m glad someone said it. | 13 |  | 4.67 |
| 6 | I have compassion for you and all other brain damage sufferers | Haha this interaction is hilarious | 17 |  | 4 |
| 7 | HOW DO YOU DROP THAT YOU IDIOT?! | Looking down field before catching it. Gotta grab that, agree! | 14 |  | Researcher- selected |
| 8 | Yeah, I hate our fanbase, so knee jerk. You all suck and need other hobbies. | You said it my brother. | 23 |  | 5 |
| Retaliatory | 9 | Do you realise how silly you sound? | Calls people silly but provides zero evidence for anything he says… Nice. | 19 | Researcher- selected |  |
| 9\* | .Aren’t you quite the smug prick? | “Well hello there pot.”  “Fuck off kettle, you black bastard.” | 15 | Researcher-selected |  |
| 10 | Because you’re a dick | So, your ignorance is caused by me? | 23 | Researcher- selected |  |
| 11 | You dumb bastard. It’s not a schooner… it’s a Sailboat | A schooner IS a sailboat, idiot! | 16 | Researcher- selected |  |
| 12 | So vicious. You slay me with your pathetic insults. | You “slay” yourself with your unmitigated ignorance. | 16 | Researcher-selected |  |

*Note*. Pair numbers with asterisks were used in the proposed experiment as replacements for those in the pilot

**Appendix B: Measures developed for Pilot and Proposed Experiment**

| Free to contribute from 0 (Not at all) to 6 (Very likely/Very willing) |
| --- |
| 1. How likely would you be to express your *true* opinion to this group? |
| 2. How likely would you be to contribute to this conversation? |
| 3. (Pilot) If you were to post to this forum, to what extent would you feel the need to hide what you really think from the rest of the group? |
| 3. (Proposed experiment) How willing would you be to join this conversation? |
| Toxicity dissuaded from -3 (Strongly disagree) to +3 (Strongly agree) |
| 1. (Pilot) The response fixes any concerns I have about the first comment.  1. (Proposed experiment) The response will encourage the first commenter to post more positively in the future. |
| 2. (Pilot) The response is an appropriate way to address the toxicity of the first comment. |
| 2. (Proposed experiment) The response will encourage the first commenter to post in a more negative tone than before.\*  3. The response will discourage the first commenter from continuing to post in the same negative tone as before. |
| 4. The response will make the first commenter reconsider what they initially posted. |
| Manipulation check questions |
| **First comment toxicity:**  (Pilot) What is your first impression of the person who made the first comment above? (-3 (Very Negative) to 0 (Neutral) to +3 (Very Positive).  (Proposed experiment) Please rate the first comment you read above on the following scale: -2 (Very toxic - a very hateful, aggressive or disrespectful comment that is very likely to make you leave a discussion), -1 (Toxic - a rude, disrespectful or unreasonable comment that is somewhat likely to make you leave a discussion), 0 (Neither), +1 (Healthy contribution - a reasonable, civil or polite contribution that is somewhat likely to make you want to continue a discussion) or +2: (Very healthy contribution - A very polite, thoughtful or helpful contribution that is very likely to make you want to continue a discussion)  **Reply benevolence, correction, retaliation**  (Proposed experiment only) Consider the replies to the toxic comments you read. Overall, to what extent did these replies… (participants responded from 0 (not at all) to 6 (extremely) |
| 1. …demonstrate benevolence (politeness, understanding, and empathy) for the initial commenter? |
| 2. …appear to correct the initial comment?  3. …appear to retaliate against the initial commenter? |

\*Reverse-scored item

**Footnotes**

1. To check whether our individual difference measures, Willingness to Self-Censor and comfort with offensive language, differed by condition, we conducted two one-way ANOVAs predicting the measures from condition. Neither was significant (both *p*-values > .59), suggesting that our manipulation did not affect these individual differences. [↑](#endnote-ref-1)
2. Inspecting the confidence intervals around the condition means, we noticed that the interval appeared wider in the retaliatory condition (95% *CI* [0.84, 2.41]) than in either the benevolent correction (95% *CI* [1.77, 2.69]) or benevolent going along (95% *CI* [1.82, 2.26]) conditions. We conducted an exploratory multilevel model predicting first impression of the toxic commenter from condition, pair (1-12) and the interaction between condition and pair, nested within participant (ICC = 0.15). The interaction was significant, *F*(6, 342) = 16.03, *p* < .001, suggesting that conversation pairs differed in first impression across condition. Inspecting a plot of the first impressions across condition, one specific pair appeared more positive than the others: pair 9 (pair 1 of the retaliatory condition; see Figure #). The specific comment (“Do you realise how silly you sound?”), on inspection, did not appear to use to be especially negative, either. Further, excluding ratings of this particular pair from the analysis predicting free to contribute resulted in a significant condition effect (*F*(2, 6.50) = 7.69, *p* = .019). Based on this, we decided to replace pair 9 in the proposed experiment. [↑](#endnote-ref-2)
3. Similar to the free to contribute analysis, we also ran the multilevel model eliminating pair 9; this did not change the effect of condition on toxicity addressed (*p* = .10). [↑](#endnote-ref-3)