Dear Alison Young Reusser,

Your article, entitled **Responding to Online Toxicity: Which Strategies Make Others Feel Freer to Contribute, Believe That Toxicity Will Decrease, and Believe that Justice Has Been Restored?**, has now been reviewed. The referees' comments and the recommender's decision are shown below. As you can see, the recommender suggests revisions.

We shall, in principle, be happy to recommend your article as soon as it has been revised in response to the points raised by the referees.

When revising your article, we remind you that:

1) Data must be available to readers, either in the text or through an open data repository such as Zenodo (free), Dryad (pay) or some other institutional repository. Data must be reusable, thus metadata or accompanying text must carefully describe the data;

2) Details on quantitative analyses (e.g., data treatment and statistical scripts in R, bioinformatic pipeline scripts, etc.) and details concerning simulations (scripts, code) must be available to readers in the text, as appendices, or through an open data repository, such as Zenodo, Dryad or some other institutional repository. The scripts or code must be carefully described so that they can be reused;

3) Details on experimental procedures must be available to readers in the text or as appendices;

4) Authors must have no financial conflict of interest relating to the article. The article must contain a "Conflict of interest disclosure" paragraph before the reference section containing this sentence: "The authors of this article declare that they have no financial conflict of interest with the content of this article.";

5) This disclosure has to be completed by a sentence indicating, if appropriate, that some of the authors are PCI recommenders: "X is a recommender at PCI Registered Reports.".

When your revised article is ready, please:

1) Upload the new version of your manuscript onto your favorite open archive;

2) Follow this link https://rr.peercommunityin.org/user/my_articles or log onto the PCI Registered Reports website and go to 'For Contributors -> Your submitted preprints' in the top menu;

3) Make your changes to the title, summary, link to the article (or its DOI) and keywords if necessary by clicking on the 'Edit Article' button. If you haven't already, you may also upload a picture or an illustration for which you own the rights – this picture will be used to illustrate your article, if recommended;

4) Write, copy/paste or upload (as a PDF file) your point-by-point response to all comments (outline each change made in your manuscript or providing a suitable rebuttal reply to the recommender's and reviewers' comments) by clicking on the 'Write, edit or upload your reply to recommender' button. You must also upload (as a PDF file) a revised version of your article, with all modifications indicated in Track Changes mode;

5) When you are ready to submit your new version, click on the 'Save & submit your reply' button.

Once the recommender has read the revised version, they may decide to recommend it directly, in which case the editorial correspondence (reviews, recommender's decisions, authors' replies) and a recommendation text will be published by PCI Registered Reports under the license CC-BY.

Alternatively, other rounds of reviews may be needed before the recommender reaches a favorable conclusion. They may also reject your article, in which case the reviews and decision will be sent to you, but they will not be published or publicly released by PCI Registered Reports. They will be safely stored in our database, to which only the Managing Board has access. You will be notified by e-mail at each stage in the procedure.

We thank you in advance for submitting your revised version.

Yours sincerely,

The Managing Board of PCI Registered Reports

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 Reusser, Kristian M. Veit, Elizabeth A. Gassin, and Jonathan P. Case

https://osf.io/hnds5?view_only=2b45b35cf37e46e5818a40bf79fc981d version 1
Submitted by Alison Young Reusser 02 Aug 2023 05:30

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 pre-registered experiment (n = 1357 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 benevolently correcting the toxicity was seen as the most helpful option for all three dependent measures. These findings suggest that treating toxic commenters with empathy, understanding, and politeness while correcting their toxicity can be a useful strategy for online bystanders who want to intervene to improve the health of online discourse. Preregistered Stage 1 protocol: https://osf.io/hfjnb (date of in-principle acceptance: 01/23/2023).

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

Round #1

by Chris Chambers, 08 Sep 2023 14:45
Manuscript: **https://osf.io/hnds5?view_only=2b45b35cf37e46e5818a40bf79fc981d** version 1

Reviews

Reviewed by Corina Logan, 04 Aug 2023 12:02

Dear authors,

Congratulations on completing your study! The results are really interesting and will be useful for people when navigating online interactions. You did such a good job with your Stage 1 that it was very straightforward to review your Stage 2 using the track changes document. Thank you for such clarity!

My comments are as follows (using page numbers from the track changes .docx file)…

Abstract - “pre-registered experiment” and “Preregistered Stage 1 protocol” - a preregistration is different from a registered report, so please change the language to registered report.

> Thank you – I used the language recommended on the Guide for Authors about halfway down on this page:
>
> “Note that the abstract of the Stage 2 manuscript must contain a direct URL to the preregistered Stage 1 protocol and state the date of preregistration, e.g. “Preregistered Stage 1 protocol: URL (date of in-principle acceptance: DD/MM/YYYY)”.

Abstract - “benevolently correcting the toxicity was seen as the most helpful option for all three dependent measures” - the word “for” makes me think that the dependent measures were something other than benevolently correcting, etc. If so, please clarify what the dependent measures were. If not, you can replace “for” with “of”.

> Done!

Table 1 - you could add a column that states the results. If you do, perhaps it would be more useful to move the table to the results section. I think this would be useful because there are so many results and they are hard to keep track of, so having them in one place would be really handy.

Thank you for this suggestion – I've left the main Study Design Table as is, but added a table (Table 6; page 39) which summarizes the results in line with each hypothesis from the initial table. This helped us see that we had mistakenly forgotten to include a manipulation check analysis comparing the three conditions on how retaliatory the replies appeared to be. This has been added on page 43.

We also realized a mistake which has resulted in many edits to the multilevel analyses in both the main and pilot experiments. The changes as a result of these edits are summarized below. **We understand that typically, editing information that had already been approved in Stage 1 is not allowed, but since this represents an error in the pilot analyses which we have now fixed we hope this is an exception to that rule.**

Explanation: in our multilevel analyses (predicting free to contribute and toxicity dissuaded) we had intended to run (quoting from p. 35): "multilevel regression models nesting ratings within pair (1-12) and participant predicting ratings of each separate conversation." We mistakenly failed to include conversation pair as a nesting variable **in both the pilot and the main experiment**, so we have added pair as a nesting variable in the analyses for:

- First Impression of the Toxic Commenter
    - PILOT (p. 21): Condition effect was non-significant and remains non-significant. We have removed first impression of the comment toxicity from the free to contribute and toxicity dissuaded analyses*
    - MAIN (p. 36): Condition effect had been significant but is not significant now. We have removed first impression of the comment toxicity from the free to contribute and toxicity dissuaded analyses.
- Free to Contribute
    - PILOT (p. 21-23): Condition effect remains non-significant. Effect sizes are weaker.** Figure 2 updated.
    - MAIN (p. 36-37): Condition effect remains significant. Benevolent going along is no longer lower than benevolent correction. Benevolent correction is still higher than retaliatory. Hypothesis 1 still disconfirmed.
- Toxicity Dissuaded
    - PILOT (p. 23-25): Condition effect remains non-significant. The comparison between the benevolent correction and benevolent going-along conditions was not significant but is so now. Effect sizes are similar. Figure 3 updated.
    - MAIN (p. 38-40): Condition effect remains significant. Benevolent going-along remains lower than benevolent correction. Benevolent going-along no longer lower than retaliatory (no difference now). Correction still higher than retaliatory. Effect sizes similar. Hypotheses 2a now disconfirmed, but 2b still supported.
- Word Count
    - This also changed the word count analyses (see supplemental materials):

- Free to Contribute – condition difference maintained, benevolent going along no longer lower than benevolent correction (similar to fixed analysis above). Other differences maintained. Hypothesis 1 still disconfirmed. Word count no longer related to free to contribute.
- Toxicity Dissuaded – condition difference maintained, benevolent going along still lower than benevolent correction, benevolent going along no longer lower than retaliatory (no difference now, similar to fixed analysis above), correction still higher than retaliatory.

*PILOT: In the previous version of this analysis, the comparison between the Retaliatory and Benevolent Correction conditions had confidence intervals which barely overlapped and an associated p-value of .06, leading us to control for comment toxicity in our pilot analyses. After fixing the nesting mistake, the confidence intervals now substantially overlapped and the p-value was much larger (.49), so it no longer made sense to control for comment toxicity. **We therefore removed "initial comment toxicity" from all later pilot analyses. These edits have been made to the analysis on p. 21.**

**Because the pilot effect sizes are now different, the section (p. 28) on the Smallest Effect Size of Interest wrongly states the range of effect sizes for free to contribute and toxicity dissuaded. We thought it would be misleading to just change those ranges and have instead opted to note that at the time, "we mistakenly believed" these were the effect size ranges. We've added footnotes to clarify this for the reader.

p28 - your planned sample size was 1122 participants and you mention "several hundred additional people completed the study, resulting in a sample of 1360". Several hundred people implies that 500+ additional people participated. Did hundreds of people fall out of the sample or were only 238 additional people tested?

I've replaced the confusing phrase "several hundred" with the precise number – 238.

p32 - I'm finding it difficult to follow the reasoning starting with "Using Rosseel's (2012) lavaan R package in jamovi". I'm not clear why "measuring freedom to contribute and toxicity dissuaded" needed to be combined "into a single confirmatory factor" - doesn't that combine these initially separate dependent variables into one? If so, how can they be thought of as separate variables after this? What is Item 2 and what does it mean to have removed it from this analysis? Adding a summary sentence in this paragraph or the next could help clarify whether you found "evidence of unidimensionality" (is unidimensionality the same as strongly loading onto one factor?) and for which dependent variables (it looks like both of them?).

I've removed the word "single" – the intent was to say that we conducted one factor analysis with two factors specified but I see why that was confusing.

I've added the wording of Item 2 to the text – the purpose of dropping the item was to increase the reliability of the measure and since it did not load strongly onto the "toxicity dissuaded" factor. I've added a summary sentence to the end of the paragraph beginning "We entered the next seven items" (p. 34) emphasizing that we have evidence of unidimensionality for our three dependent measures.

p35 First impression of toxic commenter - "While the initial comments did not differ between the two benevolent conditions (p = .70), those in the Retaliatory condition were rated as less toxic". Does this mean that it was the same set of initial comments (across both benevolent conditions and the retaliatory condition) that participants were reading, and that the participants in the retaliatory condition rated these same comments as less toxic? I'm having a hard time figuring out why this would be the case. Perhaps add an example here or provide a bit more clarification to help the reader follow?

This is now slightly different given the redone analyses, but I've added a clarification that "the initial comments did not differ **in perceived toxicity"** which should hopefully make that sentence less confusing. It isn't that the comments themselves were the same set, but that the ratings of perceived toxicity did not statistically change from condition to condition. This is in the section "First Impression of Toxic Commenter" on page 36.

p36 - "However, inconsistent with Hypothesis 1, participants did not feel less free to contribute in the benevolent going-along condition than the retaliatory condition; the opposite was found". The double negatives and reference to the opposite was confusing me and I had to go back to the study design table to figure out what the original prediction was. It would be easier if you rephrased to "However, inconsistent with Hypothesis 1, participants felt more free to contribute in the benevolent going-along than the retaliatory condition". Along the same lines, I would add ", rather than there being no difference between these conditions as we initially predicted" to the end of "Also inconsistent with Hypothesis 1, participants felt more free to contribute in the benevolent correction condition than in the retaliatory condition according to a post hoc comparison".

This was a bit different after the redone analyses, but I've changed to wording to avoid double negatives as you say: "participants' freedom to contribute did not differ between…" (p. 37)

I added the sentence "We had expected these two conditions not to differ" after the sentence beginning"Also inconsistent with Hypothesis 1.."

p38 - "According to a post hoc comparison, it was also significantly, though only slightly, lower than the retaliatory condition mean". Please add clarifications to this sentence similar to how I did above by explicitly reminding the reader of what the hypothesis states and how the results are consistent or not.

I've added clarification that this is inconsistent with Hypothesis 2a. I've also added a summary sentence that Hypothesis 2a was disconfirmed but 2b was supported (p. 39) and referred people to the new Table 6 which summarizes all predictions/outcomes.

Given that Hypothesis 2a was disconfirmed, I've added a brief paragraph to the discussion ("Retaliating Not Seen As An Effective Deterrent") exploring why. – p. 46

I’ve also added a paragraph to the discussion highlighting the limitation that we measured expectations of future events with “free to contribute” and “toxicity dissuaded” – we didn’t actually follow up and see whether people contributed more or if toxic posters posted less toxic content. (p. 49).

Figures - jittering the blue dots would allow the whole data set to be seen. As it is now, the blue dots cover the whole range of y-axis values and it is impossible to judge from this presentation where one would expect the mean to be.

Thank you – we created new figures 6 and 7 that show the jittering of the datapoints with darkness of the color indicating density. We have also overlaid a violin plot to clarify the distribution in each condition. Given the fact that we had to re-run the pilot multilevel analyses, as well, we replaced the pilot figures (2 and 3) with similar plots to be consistent.

p42 - “Note that these effect sizes were all small or less than small.” - please explain what this means for interpreting the results…should they have less confidence in them until future studies can replicate the findings? Also, did you have the power to detect differences in these small effect sizes? If so, then it shouldn’t matter that they are so small and you can add a note about this here. I tried to figure this out from the manuscript, but it is difficult.

Thank you – I’ve added clarifying language that the improvement in freedom to contribute between the two conditions is less than a scale point and that both means are still below the scale midpoint. This manipulation did not have a huge impact on people. It had an impact, but other manipulations should probably be tried.

In the Results > Free to Contribute section where the results for Hypothesis 1 are, there is only 1 F-statistic and it is 17.84. The other two analyses for Hypothesis 1 don’t have F-statistics so how are you determining whether you were able to detect differences? It would be good to explicitly state throughout your Results sections whether you were able to detect each result given the sampling plan detection levels in Table 1.

As part of the multilevel regression analyses, the overall condition effect is associated with an F-ratio and p-value but the specific pairwise comparisons between condition means are tested using a t-test. The t-scores and associated p-values are therefore included as tests of these differences.

p42 - “Replies which benevolently corrected the toxicity resulted in a mean near the scale midpoint” - please clarify what variable the “mean” was of (i.e., perception that toxicity had been dissuaded).

Thank you – I’ve clarified that this is a “mean toxicity dissuaded rating.”

p44 - “Benevolent Corrections as Injunctive Norms” - Something to consider adding to the discussion… The fact that the benevolent correcting strategy has been shown to be more effective, but also used less indicates that there must be some external pressure causing the reduction in the use of this strategy. One reason people might not use this strategy is that they aren't aware that it is more effective. So a lack of awareness of what works, along with some external pressure could also explain why this strategy isn't used as often.

I appreciate this line of thinking. I'm not sure it completely fits here only because we have evidence that our participants, on average at least, *are* aware (at least believe) that benevolent corrections will be more effective at reducing toxicity than the other options.

In the PCI RR questionnaire, you state that you provided the analysis code and pointed readers to this URL: https://osf.io/6vkqz/?view_only=2b45b35cf37e46e5818a40bf79fc981d. I went to the URL, but couldn't find any code. Please add the code to the OSF repo and provide a direct link to this document so people don't have to click through every file to try to find it.

I placed the jamovi files (updated with the above fixed analyses) in the OSF repository. These have all the analyses in the files themselves. I've also added a folder for R code used for multiple imputation and to generate the graphs. I don't have access to edit the PCI RR questionnaire to replace the link, but I have edited the names of many of the subfolders on the OSF page to make it clearer where the code is located.

All my best,

Corina Logan

Reviewed by Marcel Martončik

I would like to commend the authors for their very clear and comprehensive writing of all Stage 2 sections of the manuscript. Structuring them into three units (freedom to contribute, belief about toxicity reduction, and impression that justice has been restored) greatly enhances clarity.
I also want to acknowledge the choice of the research question, which is both timely and of great practical importance - gaining insight into the best ways to respond to online toxicity. I am convinced that this study will make a valuable contribution to the ongoing debate on combating hate speech and online toxicity in general.
I have no significant comments on the manuscript; below, I offer some suggestions on how, from my perspective, the informativeness and clarity of the manuscript could be further enhanced.

1) The manuscript contains a lot of descriptive statistics spread throughout the Results section, and I have found myself looking for these descriptives in the text several times. Having all the information (e.g., M, SD, range…) for all variables and conditions contained in one table would make it clearer and much easier to find.

Great idea. I've added Table 5 which summarizes the means and standard errors per condition for each dependent measure with superscript letters specifying which differ statistically and which do not.

2) Out of curiosity, I would be interested to know how many participants failed at least one attention check. I'm wondering if a relatively small incentive may potentially lead to careless

responding.
Good question. I’ve added this information in Footnote 6.

3) Within the section titled "Scale Reliability, Unidimensionality and Composites," where the authors report the results of the CFA, it would be useful to include also the results of the chi-square test, its associated df, and p-value.

I’ve included this information in addition to a footnote (footnote 8) arguing from Babyak & Green (2010) that a significant chi-square value given our large sample size is not necessarily an indication of poor model fit.

4) I found it difficult to differentiate between the three conditions in Figures 6 and 7. If it were possible to increase the scale of the figures to show more units on the scale (e.g., showing more intervals like 0, 0.5, 1, 1.5, 2 instead of -2, 0, 2), it could make the differences easier to notice.

Agreed. I tried creating tick marks every half-scale point but it was visually confusing; Even tick marks at 0, 1, 2, etc. seemed crowded. The new violin plots we’ve produced make the differences between the group means clearer, though, and each interval (0, 1, 2, etc.) is marked with a horizontal line.

5) If a simpler explanation of the meaning of the effect sizes, particularly for H1 and H3, could be provided, it would enhance the understanding of the results. For instance, how much freer to contribute does d = 0.15 means? Perhaps offering an interpretation of unstandardized coefficients could assist in achieving this clarity.

I’ve added an explanation in the Main Experiment Discussion under Hypothesis 1: Free to Contribute that the difference is two-thirds of a scale point on a 0 to 6 scale, which will hopefully clarifying the meaning of the effect size. I’ve done something similar for the section on Hypotheses 3a and 3b.

6) In the Discussion section under the heading "Benevolent Corrections as Injunctive Norms," the authors correctly conclude that the effectiveness of this type of comments was not corroborated or that participants "neither agreed nor disagreed that benevolent corrections would lead to less toxicity from that commenter." For this reason, I find the results more relevant for discussing the role of retaliatory and benevolent going along within the Focus Theory of Normative Conduct, as suggested in the Introduction. Therefore, it seems to me that focus of the discussion could be put more on explaining the roles of retaliatory and benevolent going along behaviors that participants perceived as ineffective, rather than on the benevolent correction behavior, where the results were inconclusive (The mean for the benevolent correction condition was -0.02 on a Likert-style scale).

Agreed – I’ve added a brief discussion (“Retaliating Not Seen as an Effective Deterrent”) which brings up the retaliatory vs. going-along conditions not fitting with our Focus Theory prediction, and our attempt to explain why that might have happened.

7) After initially reading the Main experiment discussion, I had several questions that I only found answers to after reading the General discussion. Merging both discussions (main experiment discussion and general discussion) into one coherent section could potentially provide a smoother reading experience, at least from my perspective.

To keep the discussion of the pilot and the main experiment as parallel as possible, I'd like to preserve the structure as it is, but I appreciate that it means readers don't get all the information at once.

2A. Whether the data are able to test the authors' proposed hypotheses (or answer the proposed research question) by passing the approved outcome-neutral criteria, such as absence of floor and ceiling effects or success of positive controls or other quality checks.
Data from manipulation checks supported the claim that the authors have successfully manipulated how benevolent and how correcting the conversations were.
To make data anonymous, I would recommend removing column „workerId" from csv data files.
Done.

2B. Whether the introduction, rationale and stated hypotheses (where applicable) are the same as the approved Stage 1 submission.
The authors have remained consistent in their framing of the study at Stage 2.

2C. Whether the authors adhered precisely to the registered study procedures.
All deviation from the original protocol are shared within the manuscript and are reasonably justified.

2D. Where applicable, whether any unregistered exploratory analyses are justified, methodologically sound, and informative.
The manuscript does not contain additional exploratory analyses.
We did include an exploratory word count analysis but it is labelled as such and does not meaningfully change the main findings.

2E. Whether the authors' conclusions are justified given the evidence.
In discussing the results, the authors strictly adhere to the data obtained.