Dear Managing Board,

Thank you very much to Corina Logan and Marcel Martončik for their substantive feedback on our initial manuscript; their contributions have resulted in improvements to the paper's theoretical foundations, overall flow, and precision of hypotheses. The proposed experiment has also been improved methodologically as a result of their suggestions. We have uploaded both a version of the edited manuscript with changes tracked and a version with those changes accepted. A point-by-point response to the reviewers' comments is below.

Thank you again to everyone involved with the PCI Registered Reports process.

Alison Young Reusser
Houghton University

Point-by-point replies to reviewer comments (replies in blue):

REVIEW 1

Major Revision

I have now received two very detailed and constructive reviews of your submission. As you will see, the comments are extensive and identify a range of areas requiring careful revision in order to satisfy the Stage 1 criteria. Without providing an exhaustive overview, the main issues to address across both reviews are: the clarity of the research questions including the strength and clarity of the theoretical framing, the sample size justification, the precision of predictions and in particular the precision of the contingent interpretrations given different outcomes, clarifying the definition of constructs and measurements, and consideration (and clarification) of the manipulation checks. The reviewers also offer valuable suggestions for improving the clarity and structure of the presentation in key places.

Overall, based on my own reading I think your submission is promising, and if you can provide a comprehensive revision and response that satisfies the reviewers, then I believe your manuscript will eventually be suitable for Stage 1 acceptance.

Reviews

Reviewed by Corina Logan, 05 Aug 2022 21:58

This Stage 1 Registered Report (RR) aims to test three hypotheses about how free participants feel in contributing to online conversations with toxic comments, and whether participants feel a specific toxic comment or situation has been addressed and resolved by a given response to that comment.

I applaud the authors for fleshing out predictions for multiple possibilities of the outcomes - it is such a great way to a priori consider how you will interpret whichever result ends up being supported and to make these alternatives part of the whole research program (rather than just discussing a favorite prediction, which might not be supported).

The RR is well developed and carefully thought out. Please see my comments below (minor and major mixed together, following the page numbers of the RR) for ways in which I think it could be clearer and for a couple of (surmountable) issues.

1. Abstract - state what the n=126 and n=800 refers to…the number of Reddit conversations? Or comment-reply pairs? Or people?
   This has been clarified in the abstract.

2. Page 3, par 1, sentence 1: perhaps start with a broader sentence to introduce the idea for your article and why this topic matters. Starting with the big problem that you are aiming to solve could be a good angle. And then it would make sense why you are jumping in to using Google's codebook, definitions, and why it matters how people respond to toxic posts. Explain what API stands for.

   Thank you. To address this, I've reworked the first part of the introduction as follows:
   - I've dropped the reference to the acronym API since that would require defining computer-science-related concepts and I don't want to get too into the weeds. The main point is that Google provides tools to classify the toxicity of text and this is how they define something as toxic.
   - I've also brought in some more specific discussion of our work on Reddit conversations, explaining how we define a benevolent reply and how often it occurred in our previously-published data (Young Reusser et al., 2021)
     - This also allows for an earlier definition of Benevolent Correction vs. Benevolent Going-Along, and justifies how we got those two categories of replies
     - I've moved the lion-share of these details to supplemental materials to streamline the rest of the paper
   - I've emphasized the problem we're trying to solve more:
     - Online toxicity is common
     - Some propose top-down solutions, like banning
     - We propose bottom-up solutions, like direct user replies, might help, but we want to see what sorts of replies are best

3. Page 3 "While Kolhatkar and Taboada (2017) have argued that comment toxicity is unrelated to its ability to promote civil" - clarify what "its" refers to. Reddit? And clarify whether you think that responses to news articles will be different from interpersonal interactions. As a reader, I don't know how to interpret this sentence as it relates to your research - does this study have an impact on the interpretation of your results? Or are news articles a different context and you think the responses there won't be relevant to your context?

   I have clarified "its" as follows: "comment toxicity is unrelated to that comment's ability to promote civil…"

   And yes, I do think that responses to news articles will be different from interpersonal interactions and have tried to clarify that point.

4. Page 4 - "one-on-one conversation can persuade the original commenter to change their views" - in what context? Change views about beliefs or change views about participating in

an online conversation? It seems like the former because I assume that the one on one conversation happens in person? If that is the case, it would be good to make an argument about whether in person interactions apply to online interactions to predict whether this finding would apply to your research question's online context.

I've moved this citation (Wright et al., 2017) to the section on the second dependent variable (whether toxicity has been reduced) because it makes more sense there. I've clarified the context of that research (Twitter). I've also added another citation to bolster that section (Hangartner et al., 2021).

5. Page 5 - "Are there any differences among them in how free participants feel to participate?" - differences among what? The three strategies you outlined in the previous sentence?

   I've deleted the two paragraphs that include this sentence. Since I've inserted a discussion of the Young Reusser et al. (2021) Reddit dataset & define our formulation of benevolence earlier in the manuscript, these paragraphs are no longer necessary.

6. Page 5 - "Perhaps benevolent correction of the toxicity is the best strategy" - the best strategy for what and in what context? I can imagine that the best strategy could differ depending to the goals/motivations of the forum/commenter/observer and whether repeated interactions were required with these individuals in the future.

   This sentence has been removed.

7. Page 6, Hypothesis 1a - how are "benevolent replies" different from 1b "benevolent corrections"? It seems like the latter would be a sub category of the former, but it just depends on how you categorized each term and whether there is overlap in the data that will be used to evaluate each hypothesis (i.e., all of the data from 1b is included in the 1a analysis). This becomes clear later in the RR, but I think it would be good to mention here near the beginning for clarity.

   Based on some additional theory work (Spiral of Silence) I've added because of another reviewer's feedback, I've reduced the first hypothesis down to a single statement, so this is no longer as confusingly worded.

8. Page 7 - "had more respect for the second person if they condemned vs. empathized with the target". I'm not clear on which condition elicited more respect for the target: if the observer had an attitude toward the target that was condemning or if they empathized with the target. Could you provide more detail?

   I tried rephrasing this to make it clearer.

9. Study design table > Interpretation given different outcomes: how will you determine whether or not there is a difference between the means?

   I've included information that comparisons will be judged at the .05 level.

10. Study design table > Q2 > rightmost column: replace "I" with "it" in "If H2a is supported, I…"

    Thank you - done

11. Study design table > Manipulation check - correcting > Hypothesis - should retaliatory be added to this cell? It looks like it because the retaliatory condition is in the ANOVA and in the interpretation.

    Thank you – I've tried to clarify this. We aren't so much concerned that the retaliatory condition is rated as more corrective than the benevolent correction condition – as long as benevolent corrections are rated as more correcting than benevolent going-along, we can argue that even though both kinds of replies are benevolent, one is more corrective and the other is less so.

12. Study design table > Manipulation check - toxicity > Hypothesis - "Ensure the first impression of each toxic commenter is similar across conditions." The first impression of the participant as they participate in the experiment? Or the first impression of the experimenters who are categorizing the comments as toxic, benevolent, etc? Again, this becomes clear later on, but good to mention early in the RR to help readers follow.

    Thank you – I've clarified this

13. Page 11 - for the interrater results, please state what test was used.

    I've moved this section to the Supplemental Materials, but have clarified that I used Cronbach's alpha

14. Page 11 - "The research assistants also re-rated the toxicity of each initial comment" - will you clarify how the toxic comments were classified as you did for the benevolent comments? Was a comment classified as toxic if it received a 1 or less on the benevolence scale? Or did toxic comments have their own scale? A few more details would be helpful here.

    I've added clarifying info in the supplemental materials to address this.

15. Figure 1 legend - please explain the x and y axes here, the sample sizes for each panel, what each dot represents, and what the violin shape represents. Also, a summary of the take home message would be useful. Do you need to cite the data you used here or is the data unpublished?

    I've moved this to the supplemental materials, but the number of ratings (613) is included in the figure note. I've also explained the violin plot in more detail and included a take-home message. The width of the violin indicates the frequency of the response. This data is unpublished.

16. Page 12 - "A pdf of our Qualtrics survey and deidentified pilot data can be found…" Please indicate the file name so readers know where to find this data. I didn't see a pdf of the Qualtrics survey at the OSF project.

    That file is now available and I've specified the filename.

17. Page 14, top par - how were the researcher-selected replies rated on the benevolence/retaliatory scales? If they weren't rated, then why were these treated differently and how were they categorized?

I've clarified that retaliatory replies were selected from the replies that scored the lowest in benevolence in addition to the requirement that they also had to be negative, aggressive, dismissive, and/or rude.

However, to bolster our case, I've added a manipulation check question to the proposed experiment – participants will be asked to report the extent to which the replies appear to retaliate against the initial commenter from 0 (not at all) to 6 (extremely).

18. Page 14, 2nd par - just to clarify, a "conversation pair" is a comment-reply pair? It would be good to either make sure this is clear throughout or change the term to something more intuitive.

    I've changed the "conversation pair" reference to "comment-reply pair" throughout the paper.

19. Pilot study - throughout this section there are alphas reported, however it is not clear what they refer to - interrater reliability of a particular interpretation of, for example, the toxicity of the initial comment? Please clarify throughout and include the name of the test and a description of what the statistic represents.

    I've clarified that these are Cronbach's alphas. This is a commonly reported statistic in psychology research measuring the internal consistency or reliability of a multi-item scale.

20. Page 16 - "Social media use was included to describe our sample" How does social media use describe your sample?

    I've tried to clarify this. It is included to help characterize the extent to which the people in our sample are familiar/unfamiliar with online conversation. It is descriptive, though, because we don't intend to use it in any inferential analyses.

21. Page 16 - should you list your IRB protocol number? I'm not sure how it works with studies on humans, but studies on non-humans have to list this in all articles.

    This is not usually a requirement for human-participants research (I've never run across it, at least).

22. Page 16 - please clarify that pair 1-12 means 4 comment-reply pairs multiplied by 3 conditions.

    Thank you – I've clarified this.

23. Page 16, last par - please show the data from the other benevolent condition as well so readers can evaluate what a "marginal" difference is.

    I've included the benevolent going-along condition's mean as well as the planned comparison between benevolent going-along and benevolent correction.

24. Page 16, last par - "The effect of condition was not significant, however, given that the difference between the retaliatory and benevolent correction conditions was marginal (planned comparison $t(114) = -1.89$, $p = .061$), we decided to control for the first impression in all multilevel analyses" It looks like the "marginal" difference was determined based on $p=0.061$? If that is the case, what was your preplanned cut off for determining whether there

was a difference between conditions/means/etc or not? If the cut off was p=0.05, then there is no "marginal". It is either on one side of the threshold or not (see references below for further discussion on this topic). I realize this was for your pilot study and not your proposed study, however your decision to include first impression as a fixed effect in the analyses for the proposed study is likely based on this finding. If this is the case, because of your non-significant finding in the pilot study, the first impression should be removed from the proposed analyses.

I've taken out the "marginal significance" language. My concern is that since the pilot is underpowered, a p-value of .06 suggests that with sufficient power, the retaliatory condition's initial comments might be rated as more toxic than the other two conditions, meaning we should control for perceived toxicity (the variable we are replacing "first impression" with). See Figure 1 in this paper by (https://arxiv.org/pdf/1311.0081.pdf) for an argument from Monte Carlo simulations that smaller p-values are more commonly observed where the null hypothesis is false.
James C Boyd, Thomas M Annesley, To *P* or Not to *P*: That Is the Question, *Clinical Chemistry*, Volume 60, Issue 7, 1 July 2014, Pages 909–910, https://doi.org/10.1373/clinchem.2014.226282

See also Amrhein et al. (2017)'s argument against dichotomous use of p-values: "Consistent with the recommendations of the late Ronald Fisher, *p*-values should be interpreted as graded measures of the strength of evidence against the null hypothesis (abstract, Amrhein et al., 2017).

Amrhein V, Korner-Nievergelt F, Roth T. 2017. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. PeerJ 5:e3544 https://doi.org/10.7717/peerj.3544

I've added an argument in the pilot about the confidence intervals only barely overlapping for the retaliatory and benevolent correction conditions, again hoping to justify controlling for first impression in the pilot. Note that I am not trying to argue from the pilot that we should necessarily control for perceived toxicity in the proposed study- I plan to only control for perceived toxicity in the proposed study if (given the much larger proposed sample size) it differs by condition at the .05 level.

25. Figures 3 & 4 legends - please clarify what the circles refer to - the means?

    Done.

26. Pages 18-20 - "without covariates" is mentioned a few times, but I'm not sure what this means when the analyses were run with the covariates.

    I clarified this right before the section "Free to Contribute" in the pilot analyses and at the beginning of the proposed analyses for the proposed experiment. Conducting the analysis both with and then again without covariates is recommended by Segerstrom (2019) – see the paper here: https://link.springer.com/article/10.1007/s12529-019-09811-5

    Citation: Segerstrom, S.C. Statistical Guideline #3: Designate and Justify Covariates A Priori, and Report Results With and Without Covariates. *Int.J. Behav. Med.* **26**, 577–579 (2019). https://doi.org/10.1007/s12529-019-09811-5

27. Page 19 - “Comfort with offensive language was not related to toxicity addressed, p = .23” Please add the rest of the test statistics here as in the other sentences.

    Done.

28. Legends for Figures 4 & 5 and those thereafter as well - please add how to interpret the y axis. Do negative numbers mean participants felt like the toxicity was made worse, 0 = toxicity not addressed, and positive = toxicity addressed?

    Done.

29. Given that the pilot study found no significant correlations for 2 of the 3 hypotheses, it might be a good idea to add to the study design table in the Interpretation column how you will interpret when there is no correlation and what theory this would contradict.

    Done.

30. Also, was the pilot conducted according to the hypotheses in this RR? It would be good to note what the pilot study hypotheses were at the beginning of its section.

    I’ve specified that the pilot uses the same hypotheses as the proposed experiment.

31. Page 21, pars 1 & 2 - please omit the sentences that mention “weak evidence” and “marginal” - these were not statistically significant, which is the measure you chose to determine whether there were differences or not (see my comment above and references below).

    Done.

32. Page 24 and throughout - “This suggests that the manipulation of how benevolent and how correcting the Reddit conversations were was/was not successful” According to how I understand the experiment, I think you categorized the responses and not that you manipulated the responses or the participants. When I think of a manipulation, I think of designing the experiment such that the behavior of the participants changes across the study because of the experiment. If you agree, I would replace the term manipulation with categorization or something similar.

    We randomly assigned participants in the pilot (and plan to randomly assign participants in the proposed experiment) to read one of three types of replies (benevolent corrections vs. benevolent going-along vs. retaliations), so this fits the definition of an experimental manipulation.

33. Page 25 - explain what ICC is and how to interpret it on first mention.

    I’ve added an explanation the first time this is mentioned.

34. Page 26, last sentence - there is only a p value place holder; please add the rest of the statistics as in the rest of the paragraph.

    Done. Thank you.

35. Page 32 - did all co-authors approve the submitted version for publication? It looks like only one author did, however all authors need to approve of articles submitted on their behalf.

All coauthors have now approved the submitted version.

Throughout:

36. - the terms “benevolent going-along” and “benevolent endorsement” terms are used in different places. I would choose one term and stick with it to avoid confusion.

    Fixed.

37. - the axis labels look like they are the raw variable names and would be clearer if they were relabeled to assist readers with interpretation.

    Thank you – this has been fixed.

>Assessing the RR according to PCI RR’s Stage 1 criteria:

>1A. The scientific validity of the research question(s).

The research questions are scientifically valid.

>1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable.

The proposed hypotheses are logical, rational, and plausible, and I suggested adding interpretations for the possibility that there are no correlations (see above).

>1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).

The methodology and analyses are feasible and I suggested a change to improve the soundness (see the comment on marginal significance above).

>1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.

The methodological detail is clear and replicable. I had a suggestion regarding the analysis pipeline to further reduce analytical flexibility (see above regarding marginal significance).

>1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).

The authors conduct categorization validation checks to ensure the three types of responses were perceived as belonging to their assigned categories.

I wish you the best of luck in conducting your study!

All my best,

Corina Logan

Max Planck Institute for Evolutionary Anthropology

References

Lybrand et al. 2021. Investigating the Misrepresentation of Statistical Significance in Empirical Articles. https://dc.etsu.edu/honors/646/

Nozzo. 12 February 2014. “Scientific method: statistical errors”. https://www.nature.com/articles/506150a

Otte et al. 2021. Almost significant: trends and P values in the use of phrases describing marginally significant results in 567,758 randomized controlled trials published between 1990 and 2020. https://doi.org/10.1101/2021.03.01.21252701

Pritschet et al. 2016. Marginally Significant Effects as Evidence for Hypotheses: Changing Attitudes Over Four Decades. https://statmodeling.stat.columbia.edu/wp-content/uploads/2016/06/Pvalues.pdf

Reviewed by Marcel Martončik, 14 Aug 2022 04:53
Download the review

REVIEW 2

I thank the authors for their submission, transparency, and adherence to open science practices. I also thank the recommender for the opportunity to review this interesting submission. I hope the following comments, suggestions, and questions help strengthen and clarify this submission.

The scientific validity of the research question(s)

1. Purpose and rationale of the study

From my point of view, the authors have chosen a topic that is both very actual and lacks accumulated knowledge. At least this is true for (related) hate speech and psychological research since this construct is rather studied by other sciences. In the first sentence of the Introduction, the rationale for conducting a such study is implicitly suggested. I would suggest authors to make it more visible and explicit. Specifically, I would suggest structuring Introduction in a way that it will more clearly formulate: 1) what is the actual problem that needs to be solved? 2) Why is this a problem (e.g., no empirical studies, mixed evidence, missing knowledge base…), and why it is important to solve this problem (e.g., Why is fairness or justice important in the context of discussion forums? or whether the toxicity has been addressed?), 3) How this RR will help to solve this problem.

I've overhauled the opening paragraphs of the introduction to highlight the actual problem, why it is a problem, and why it should be solved. While I leave the treatment of the three outcome measures (free to contribute, toxicity addressed/reduced, and overall fairness) to later sections where I can work on justifying their inclusion, I do think that this section is much more compelling than it was before. Thank you for this feedback.

What I found missing here (and I will discuss this later) is framing of the study in a broader context, e.g. theory. What kind of theory/theories discuss processes which are in the background of the studied behavior (specific replies/posts/comments and their effects…?) and could explain studied behavior (e.g. effect of retaliatory response on the increase of engagement?)?

This is very helpful.

1. I've added a section describing the Spiral of Silence Theory (Noelle-Neumann, 1977) to help frame my predictions regarding how free people feel to contribute (RQ1).
2. Cialdini et al.'s (1991) formulation of descriptive and injunctive norms provide context for RQ2 (toxicity has been addressed/reduced). This second theory has actually resulted in a modification of my hypotheses for RQ2. Given the theory, it no longer makes sense to predict in Hypothesis 2b that retaliatory replies will do a better job of addressing/dissuading toxicity than benevolent corrections. I've adjusted it to predict that *either* benevolently correcting *or* retaliating will do a better job than benevolently going along with the toxicity.
3. Wenzel and Okimoto's (2008) theory regarding the psychological motivations underlying retributive vs. restorative justice provide context for RQ3. While the hypotheses have not changed, I think this helps clarify why those hypotheses were generated. I have also changed the name for this measure from a sense that fairness has been restored to a sense that justice hsa been restored to better fit the language of the theory.

4. Formulations of RQs

It was not directly clear to me from the manuscript what is the first research question. It would certainly help to label it explicitly, e.g. RQ1 (even though I don't like it, in the context of this manuscript it might help to solve ambiguities). In the Introduction I found several sentences that looked like an RQ:

a. 1) on p. 4.: „Toxic comments may have the potential to either decrease engagement among users or lead to more toxic engagement. Can direct replies from other users counteract this negative impact?“ Counterspeech is further discussed (its effect, its nature, hostile vs empathic) suggesting that this will be one of the central constructs of the work.
b. 2) on p. 4.: „In our work, we focus on toxicity more broadly and ask what sorts of replies to toxic posts increase engagement – specifically, how free other users feel to contribute to the conversation.“
c. 3) on p. 5: „Are there any differences among them (benevolent replies) in how free participants feel to contribute?“
d. 4) on p. 4.: „How often benevolent commenters correct the initial commenter or go along with the initial commenter“
e. 5) and one implicit formulated right at the beginning of the second sentence (p. 3): „we should expect toxic posts to decrease engagement among users in a given discussion space“ with a description of the existing research, which either supports decrease of engagement, increase, or neither of the two.
However, when I came to the Study design Table 1 I found that the actual RQ1 is formulated as follows: „To what extent do benevolent corrections, benevolent non-corrections (going along with the initial comment), or retaliatory responses to toxic comments online make observers feel freer to contribute to the conversation?”

Thank you - I have placed the exact wording for the first research question in the manuscript itself.

RQ2 and RQ3 are easily recognizable but I have struggled to find their justification. Why does this need to be studied? I had also a little trouble understanding what fairness/justice means in the context of RQ3 (“What sort of reply will help observers feel that fairness has been restored following a toxic comment?”). What is justice about? Justice in what context or a justice of what? It would help me personally if it was stated what would justice look like if it was present/restored.

I think addressing the theory gaps you mentioned (described above) serve to better justify RQ2 and RQ3.
I also believe discussing Wenzel & Okimoto’s (2008) conception of retributive and restorative justice (mentioned above) helps clarify the use of justice restoration in RQ3.

The logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses)

I would suggest providing the reader as soon as possible with the information that the present research will study only reactions to 1 reply following 1 post - a very simplified snippet of real discussion forums. In reality, however, there is a mixture of different comments/posts present in each discussion thread, some toxic others not, and a number of other factors (credibility, familiarity of commenters, type and characteristics of medium, provider, design, etc.). Legitimate question can arise such as: do factors such as the number of posts/comments/replies in a thread have any influence on the present effect? or is this effect independent of that?....etc. In the hypotheses H2 and

H3 the authors use a word "replies" instead of singular reply (e.g., after benevolent replies...). In the Methods section it follows that they will examine the response of a single reply to a single post. Does this effect change its size with the number of replies? ...

Thank you – I have changed the wording of the hypotheses so it is clear only one reply is being considered.

It was not clear to me from the wording of Hypothesis 2 („„Participants will feel that the toxicity of a specific comment has been addressed after benevolent corrections compared to benevolent going-along or retaliatory replies“) what outcome is expected. Does this mean that after benevolent going-along or retaliatory replies participants 1) will not feel that the toxicity of a specific comment has been addressed or 2) will feel that the toxicity of a specific comment has been addressed to a lesser extent?
I've reworded hypotheses 2a and 2b to to clarify this.

Will participants know the formulation "the toxicity of a specific comment has been addressed" mean? (e.g. comment was highilted as toxic; or commenter was tracked down and arrested; or commenter was banned from discussion forum; or comment was deleted from a thread, etc.)

In light of this and other comments, I think the „toxicity has been addressed“ measure is fuzzy. I've dropped two of the items and added two new items so the measure is more specific and focused on participants' sense that the reply will *reduce* the toxicity of future posts from the toxic commenter. I've changed the wording of this variable in the rest of the manuscript (toxicity reduced rather than toxicity addressed), including the title.

The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable)

Justificantion of sample size

I would like to point out the justification problem with the sample size for Pilot 1. Without a justified sample size, we have no certainty that Pilot 1 was informative. And if Pilot 1 was not informative, any SESOI based on Pilot 1 would be misleading. This is the point where I return to my former comment about missing a theory. It could potentially also help to estimate the size of the effect – e.g., look for related meta-analyses, or do a systematic search for existing literature (e.g., state the searched databases, formulate the search terms, search strings in advance...check available studies, and look for effect sizes, create summarizing table, and take the most conservative estimate).

Thank you – I have included several research examples of effect sizes related to the main DVs and selected the smallest, $f = .11$, as our SESOI. This increases our proposed sample size to 1049, which goes up to 1122 when we account for a 7% attention check failure rate (from the pilot).

A power analysis statement on p. 21 is not complete. It should contain also the estimate of effect size, and alpha level (do authors plan to use any FWER?). I have found only in the Study design Table 1 that the authors have used a default *moderate* ES from GPower software. The SESOI should be well justified and the use of Cohens or similar guidelines for selecting the SESOI is not recommended. The authors state that „Since we have three separate hypotheses and corresponding analyses, as well as analyses for manipulation checks, we multiplied this suggested sample size by three, or 618 participants.“ I dont follow why it should be multiplied instead of corrected for FWER and use e.g. different alpha levels in power analysis. Also, if the authors already have pilot data they should know about the prevalence of careless responding (5%, less, more?...) – and use this estimate for oversampling. Too big oversampling (800 instead of 600 participants) could be viewed as

unethical – why administer survey to so many people if I can get informative results with fewer participants?

I've removed the statement regarding multiplying the sample size by three and have included a statement about managing FWER.

Analyses plan

My personal opinion or recommendation is that participants does not have to be dropped completely if they will omit one or a few items from the whole survey. („Participants who do not complete any of the key measures will be dropped prior to analysis"). I would rather choose some method of imputation (e.g. MI, ME, random forests...) rather than lose so many participants (and power). Their remaining answers would have been wasted…

Agreed – we don't intend to drop participants if they are only missing one or a few items. I've tried to clarify this in the manuscript.

The authors are planning to use Mturk and at the same time plan to „drop any participants who fail an attention check." I know that problem of bots is serious but still I would consider using a less conservative solution – e.g. having more than one attention check (different combination of e.g. Mahalanobis distance statistic, bogus item, instructed response item, instructional manipulation check, honeypots questions, etc.).

Using CloudResearch in concert with Mturk actually reduces the bot problem substantially – participants have already been vetted and tend to produce high-quality data. I tend not to have to drop too many people using a single attention-check question.

Disclaimer: I am not familiar with the multilevel regression model, and could not review this procedure in detail.

Results of the Pilot

The use of nonsig. p-value to justify marginality of the difference between „the retaliatory (M = -1.63, SE = 0.10) and benevolent correction (M = -1.37, SE = 0.10)" doesnt seem to be correct. To me, this difference does not look marginal. Nonsig. results may be more probably related to the low power of the design of Pilot 1 (N = 117) but say nothing about the size of the difference.

I was using the word marginal not to refer to the size of the difference but to the fact that the p-value was close to .05 (though above it). I've removed the phrase „marginally significant," though. Because the Pilot is underpowered, I wanted to be careful with this measure of the first impression of the toxicity of the commenter. I've added an argument that the confidence intervals for the two conditions only barely overlap to justify including first impression of the toxic comment as a covariate. This is a conservative choice that only applies to the pilot – since the proposed study will have sufficient power, first impression (replaced with the variable „perceived toxicity") will only serve as a covariate if the conditions differ at the .05 level.

Use of controls (e.g., willingness to self-censor, and comfort with offensive language) should be justified and explained (rationale for their use) in the text, even if at least in one or two sentences. It looks strange to me when they suddenly appear in the analysis.

We had tried to do this in the Individual Differences section of the Materials for the Pilot study. I've edited this a bit to make it clear that we are explaining why the measures are included.

Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses

I am afraid that the proposed design of the study allows for a great deal of flexibility in the interpretation of results, as the proposal lacks clear criteria for scenarios when Ha or H0 would be supported. I will explain this on an example from the Study Design Table 1 and Column Interpretation given different outcomes; (e.g., the first hypothesis - "*Support for H1a: The retaliatory condition's mean is lower than the other two. The two benevolent conditions do not differ. If the retaliatory condition's mean is similar to at least one of the benevolent condition means, this hypothesis would be disconfirmed*"). Instead of saying "mean is lower" or "mean is similar" an interval or point estimate of effect size should be proposed as an exact threshold/s – at least how big (for Ha – alternative hypothesis) or how small (for H0 – null hypothesis) should be the difference to conclude corroboration of these hypotheses. Besides nonsignificant results do not support the absence of an effect. Instead, equivalence testing should be followed (or its Bayesian alternative). The column "Theory that could be shown wrong by the outcomes" contains references to empirical studies. Instead, as I wrote above, it would be great to state the theory that is behind commenting behavior and that could be supported/disconfirmed.

Since this is an exploratory study, we don't yet have a good sense of effect sizes to expect (outside of the pilot we conducted), so I don't think specifying an interval or point estimate makes as much sense here. Yes, non-significance is not evidence of no difference, but if we have a sufficiently-powered study, it is somewhat more informative and will give us a sense of the effect sizes future experiments might expect. I have added a statement for each set of hypotheses in the study design table that we do not have predictions as to the size of any of the effects. I've checked other PCI registered reports and while some specify effect size ranges in their „interpretation given different outcomes" column, not all do. I have specified that we are using a .05 level of significance for each mean difference.

I've included information about theories that could be shown wrong by the outcomes as suggested.

Clarity of definitions of the main constructs and concepts

1. Benevolence

Since I am not a native speaker and honestly dont feel great in English, I may be wrong but I wonder whether the naming „benevolent" for any post that „demonstrates empathy for the initial commenter, understanding of the comment's content, and a polite tone" is accurate. Benevolent and empathic should denote different things. Benevolent from the definition I have found contains an expression of goodwill, doing good, kind feelings but these aspects are not descriptive of emphatic reaction – which is mainly about recognition and understanding of thoughts and feelings of another person… In the context of counterspeech, the authors use adjectives hostile/kind/empathic but here switch to use benevolent instead of empathic. Too many different terms, in my opinion, makes it difficult to understand the message of the manuscript. E.g. on p. 5, the authors introduce 3 new terms: „In brief, though, benevolently correcting, benevolently going along, and retaliating against toxic comments were common strategies. " However, none of them are defined/explained and it can be very difficult for the reader to imagine what kind of reactions/comments are defined e.g. by the naming *benevolently correcting*. What does it mean to benevolently correct the toxicity on a discussion forum? The following two cited authors in the manuscript do not provide more clarity on this issue („Are there any differences among them in how free participants feel to contribute? One possibility is that benevolence is better suited than retaliation to boost engagement.") since Bao's research deals with „prosocial" conversations (does benevolent means the same as prosocial?) and Ziegele et al study 1 (cited by the authors) is about the preference of the specific type of comments not about readers reactions to such comments (which was the aim of their study 2). I have found

reference to Neto et al, research as not corresponding with this RR since it deals with the relationship between team performance in LoL and the nature of communication between players. In summary, the important constructs of the this RR should be clearly and comprehensibly defined.

I believe the substantial reworking of the introduction, talking earlier about our formulation of benevolence, inclusion of contextualizing theories for each hypothesis, and definitions for each kind of reply, address the above concerns. While I think Neto et al. is at least loosely relevant to the point I was making in that section, I have removed it based on your feedback.

Measurement of benevolence

On p. 11 the authors mention the use of the three-item benevolence scale. Benevolence is one of the main constructs of the study, thus I consider a description of its measurement to be provided in more detail. The authors only mention that it (is, demonstrated understanding of the comment, empathy toward the commenter, and a polite tone)). Again, the question that arises for me here is why these three items constitute benevolence. How it was constructed? Are these three items in correspondence with a definition of benevolence? I was looking for its definition but what I have found contradicts the content of this three item scale (e.g., doi: 10.2307/2025781; 10.1177/0092070303254382, 10.1111/j.1467-9833.2004.00234.x, ...)

Similarly, without a clear description of what a specific scale is measuring or how the construct is defined, it is difficult to understand what some results mean, e.g. the following sentence: („the extent to which the benevolent reply corrected the toxic comment (Correcting; interrater $\alpha = .82$), went along with the toxic comment (Going Along; interrater $\alpha = .79$), or redirected the toxic comment (interrater $\alpha = .51$; dropped because of low interrater reliability“).

Defining these terms in the introduction where I discuss the creation of the Benevolence scale clarifies the above points(p.4). One of the citations you mention on the psychology of benevolence (Brandt, 1976) seems to define benevolence as an individual difference (benevolent people are motivated to help others, pleased when others succeed, and distressed when others are in need). We are defining benevolence in terms of the behavior (a benevolent reply being one that uses a polite tone, empathy toward the commenter, and understanding of the content of the comment) rather than a set of motivations. It seems reasonable to expect, though, that a benevolent person as defined by Brandt (1976) might be more likely to make a benevolent post on Reddit, but this is outside the scope of our focus in this particular experiment.

The Lee et al. Importer benevolence seems like a very specific formulation relevant to importers and suppliers (they define it as „discretionary, extracontractual helping behavior). I think it makes sense that benevolence in an online setting would look different.

Since our definition is based on an already-published conference proceeding, as well, I think using it here allows for continuity with past work.

Measurement of feeling free to contribute

The main DV was measured (and is also planned to be measured in the confirmatory part) using three items (stated on p. 14), one is present as a sample item. It would be helpful to describe more precisely how DV was measured, how this measure was constructed (why 3 items, how they were selected? Based on what?), whether it is an aggregate score or 3 items are used separately, is the scale unidimensional? and is the support of unidimensionality available? etc. Is there any support for the validity of this measurement?

The full set of items for this DV are in Appendix B. These were written by our research team with the input of three psychologists (one I/O, one social, and one clinical/developmental) and one theologian, so at the very least they have content validity. We wrote them with the intent that they

be unidimensional, though we don't have factor analysis data to back that up given our pilot sample was so small. I am also not sure how to perform a factor analysis with multilevel data. Cronbach's alpha from the pilot suggests that for the items we retained, participants responded consistently to each (internal reliability). The item we are replacing the less-reliable item with was drawn from already-published work (Hampton, Shin and Lu (2017)), which should hopefully help with validity. This is admittedly imperfect but we could not find an already-validated measure for this particular variable.

I was thinking about the relationship between a measurement of DV and the willingness to self-censor. The first comment (from Qualtrics supplement) is: „Comment: The only relation to the thread is you and your stupidity." and then questions depicting DV ("How likely would you be to contribute to this conversation?") and the rest of the variables follow. Since the topic of the discussion to which this comment should belong is unclear (the comment itself is very vague, nonspecific, and sterile), it could be (imho) difficult for a participant to express their true intention. What I am thinking about: if I don't know (at least; among other things) the topic of discussion to which this comment belongs to, how do I know if I would join in or if I would be willing to express my true views? I am trying to explain it further: if this comment belonged to a discussion with a topic of interest to me and on my favorite medium, I'd be more likely to engage in a discussion than if I didn't have that information. The context is probably very important here. What motivates people to engage in a discussion probably is not just the comments themselves but the topic of the whole discussion in which they want to express their opinion. For this reason, the stimulus material looks a bit sterile and I would be interested about the ecological validity of the study. Besides: some of the comments are very vague, without a clearly identified topic (Comment: Your edits are dumb.; Comment: I literally cannot believe you are this stupid.) but other comments clearly belongs to a specific topic (Comment: That's why I want to play a cracked version, and not pay $60 for it, idiot.; Comment: You dumb bastard. It's not a schooner... it's a Sailboat). From what I wrote above, it seems to me that their content should be uniform (e.g. all comments woithout a clearly identified topic).

Stimulus selection is honestly limited in this case to the particular dataset we had access to (Young Reusser et al., 2021), which had already been classified according to first comment toxicity and reply benevolence. The stimulus pairs are the best examples I could find in that dataset of comments that met the requirements (high or low benevolence, correcting vs. going along as coded by research assistants) but did not mention anything controversial or too specific. You are absolutely right that participants will find it difficult to make a precise judgment on whether they would contribute without much context. I think ecological validity is strengthened in one respect by the fact that these are real-world Reddit comments, but as you say weakened by the fact that they are removed from the conversation context. In selecting our conversation pairs, we attempted to balance variation in context from conversation to conversation (essentially by removing it entirely) and focus in on the thing we're really interested in – variation in the kind of reply while keeping the comment toxicity as constant as we can (all comments are among the most-toxic on Reddit in that timespan).

I think your point is important, though, and suggests that a followup study where we create experimental conversations based on Reddit examples with 1) more context and 2) better control of content/topics would be helpful. I don't think that fits the scope of this particular RR, though.

We will include a discussion of these points in our limitations section. Thank you for this feedback!

2. Another constructs that should be more clearly defined and explained are Engagement and Fairness. What engagement is about and how it will be measured? How was the scale for fariness

constructed? What was the nature of its adaptation? (the authors state that it was adapted; e.g., what was changed and how was the validity ensured?)

With the addition of theoretical context to these sections, I believe "how free people feel to engage in the conversation" has been better-defined and explained. I have also, given a discussion of Wenzel & Okimoto's (2008) theory regarding restorative and retributive justice, renamed the fairness variable to Justice Restored.

I clarified that the justice scale was mostly verbatim from the original scale, and specified the one change we made. I also included a second example item.

Overall, it would help the readability and clarity of the text if the description of the methods followed the formulation of the hypotheses...I mean: arranging them in the same order as they appear in the hypotheses. Thus, as first would be described in the methods DV followed the description of IVs in the same order in which they appear in the hypotheses. This could make the measurement section clear and the main constructs easily identifiable.

The organization of the procedure section is hard to line up exactly with the hypotheses because of the nested design – I want to clarify that certain measures were asked multiple times for each conversation pair (Per-Pair Ratings) and that others were asked once about all pairs at the very end of the survey (Overall Ratings). The headings in the results section, though, does match the hypotheses and follows the same order.

Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).

The authors present 3 manipulation checks: first for checking benevolence, second for checking correcting, and third for checking toxicity. All three are reasonable and justified but I have a comment regarding the measurement of the third, toxicity of comments. Toxicity of comments is planned to be measured as the „first impression of the toxic commenter using seven options from -3 (Very negative) to +3 (Very positive)". I feel like these are rather two distinct constructs. Toxicity of comment on one side and the perception of its author - commenter on the other side. A comment could be considered nontoxic but at the same time have negative evaluation of its author/commenter. The impression of commenter may be related to other factors (e.g. does the comment contain grammatical errors? jargon, stylistical errors, politely expresses a point of view with which other person fundamentally disagree…).

Agreed. I've changed this item to match the wording used at the beginning of the manuscript (from Perspective, 2021). This was used to create the algorithm that initially classified the Reddit comments, so it makes sense to use it to verify their toxicity here. I've added this to Appendix B, as well.

Other Comments:

- on p. 11. in the formulation of the effect size („were above a 3.5 on benevolence") I find it useful to state (also) standardized effect size.
  - This has been moved to the Supplementary Materials from the main manuscript,but 3.5 is a reference point on the 0-6 benevolence scale from another paper, not an effect size.
- I would suggest removing Figures 1, and 2...I dont think these are super-relevant to the proposed RR.

- Figure 1 has been moved to the Supplementary Materials. I do think Figure 2 (now Figure 1) is helpful in the sense that they will eventually allow readers to visually compare the two samples on social media use and comfort with offensive language

- I found the use of the word „failed“ in the sentence “A chi-square goodness-of-fit test failed to find evidence“ strange because the test has not failed. I would suggest using other words, e.g. Using the XY test we did not find … or using equivalence testing to confirm that there is no predicted difference.
  - Changed as suggested
- I think that when using this item: „How likely would you be to contribute to this conversation?“ to measure the probability of the respondent engagement in a particular discussion it would be also important to control for their interest in the topic that is discussed in the post.
  - There are only three examples where obvious topics are mentioned, one in each of the conditions. The topics I see are: video games (one of the benevolent correction examples), football (one of the benevolent going-along examples) and sailing (one of the retaliatory examples). To avoid overcomplicating analyses with too many extra covariates, I propose I complete a supplementary analysis of the experimental data where I remove these three pairs and see if the findings change.
- I would recommend considering moving the whole Pilot 1 to the appendix and describing in the manuscript only the confirmatory testing.

  I would prefer to keep the pilot where it is as it helps set the stage for the proposed experiment.