Round #2

by Chris Chambers, 02 Nov 2022 15:31
Manuscript: **https://osf.io/zux3y?view_only=2b45b35cf37e46e5818a40bf79fc981d** version v2
Minor Revision

The two reviewers from the first round kindly returned to evaluate your revised manuscript. As you will see, both are quite positive about the revision while also noting a few remaining areas that need attention. One of the major points highlighted by both reviewers is improving the clarity of the prospective intepretration given different outcomes. As regards the comments by Marcel Martončik concerning the smallest effect size of interest (SESOI), the reviewer is correct that RRs are generally most appropriate for confirmatory studies that can specify these boundaries precisely based on prior research or theory. However, as you also note, when the SESOI is unknown then there is still a place for the RR format in providing an unbiased estimate for future studies. The key in such cases is to ensure that the interpretration given different outcomes is clear (as Corina Logan notes). For example, in the first row of the design table (p11), replace "The two other conditions do not differ" with "The two other conditions do not differ significantly", and concerning: "If the Benevolently Going Along condition's mean is similar to either other condition, this hypothesis would be disconfirmed" -- what is the definition of "similar"? It is important to be as precise as possible so that all hypotheses are falsifiable with sources of potential intepretative bias closed off as much as possible.

Thank you – we have replaced "the two other conditions do not differ" with "the two other conditions do not differ significantly" as recommended. We have also replaced the phrase "is similar to" with the phrase "does not differ significantly from" in all occurrences in the design table.

I have not put a thank you acknowledgement to these two reviewers in the manuscript, but I think it is warranted given their substantial contributions to improving it. I haven't added one yet in case there are more revisions to be done, but would it be possible to add one at some point in the process?

Please also respond carefully to all other comments from the reviewers. I look forward to receiving your revised manuscript in due course.

Reviewed by Corina Logan, 19 Oct 2022 17:38

Dear Alison Young Reusser,

You did an excellent job of revising the manuscript, thank you for addressing the comments so well! It makes the manuscript much clearer. I have only a few comments on this version.

Table 1 > Q1 > Interpretation: "This hypothesis is agnostic as to the difference between the other two conditions". What would your interpretation be if the Benevolent Going Along's mean was higher than the other two conditions? An indication of how would you interpret all possible results is warranted, even if it seems like some results would be highly unlikely because this outcome is still a possibility. Specifying these in before collecting the data will make these interpretations much more robust in the event that this unlikely result occurs. The same comment applies to Q2 and Q3.

I've gone through and added interpretations for these alternatives for each hypothesis Q1-Q3.

Regarding point 11 in the author's response, if the retaliatory category is not going to be analyzed, and it looks like it won't because it isn't part of this manipulation check, then remove this data from the data set that is being analyzed and only include the two benevolent categories in the analysis.

Thank you – I've changed this analysis to an independent samples t-test comparing the Benevolent Correction and Benevolent Going-Along conditions, leaving out the Retaliatory condition entirely.

Figures 2 and 3. It seems like the blue dots are there to delineate vertical lines from the x-axis? I would eliminate them to reduce confusion that they are data points. If they represent something about the data, please state this in the legends.

Thank you – the blue dots are data points. I've made this clear in the figure notes.

Page 31 "We plan to recruit 800 participants". I believe this number is now over 1000 given the revision?

Yes, thank you for catching this – the mistake has been fixed.

Regarding comment 26 in the author's response, I think this was the text that was added for clarification?

"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."

If so, I think an explanation should be added about why with and without the covariates are being analyzed and which set of analyses should be the ones used for coming to final conclusions.

This is based on a recommendation from Segerstrom (2019), intended to reduce researcher degrees of freedom: "In a comparison of studies that reported results only with covariates versus studies that reported results without the covariates…the set of studies that reported results only with covariates lacked evidential value…That is, the pattern of results was consistent with the presence of selective reporting…" (p. 578). I've added the note "As per a recommendation from Segerstrom (2019)," to that section of the manuscript where I describe conducting the analyses with and without covariates.

Supplementary material at https://osf.io/wa8f3?view_only=2b45b35cf37e46e5818a40bf79fc981d, Figure 1. Please label the y-axes with language used throughout the article (Benevolent Correcting, etc) and remove the word Composite because it is confusing. Unless the word composite is important, in which case it should be explained in the legend.

Done – I've changed the y-axis labels to make it clear that these are average ratings of either how correcting or how much a reply is going along with the previous one. I'm avoiding using the term Benevolent here because this is coding data where raters were asked to rate the replies on how much they corrected or went along with the prior reply, not how benevolent they were.

All my best,

Corina Logan

Reviewed by Marcel Martončik, 02 Nov 2022 08:21

I would like to say that the authors did an excellent job in rewriting the Introduction section and I appreciate that they did a lot of work. In the present form, the whole introduction is clear, and the same applies to RQs, all of them are now justified and Noelle-Neumann's and Wenzel and Okimoto's theoretical suggestions help the reader to better understand the processes behind the RQs. I also appreciate clarification of the main constructs, which helps to understand the way they were operationalized. I would also like to thank the authors for all clarifications, explanations, and edits.
Below are just a few comments on topics that I think are important to address before the study is conducted.

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

- This is just feedback for possible deeper thinking (I am not saying that the operationalization is wrong) about the operationalization of the Toxicity dissuaded (Appendix B). I have a feeling that the item "The response will make the first commenter reconsider what they initially posted." may measure something else (reconsideration) than the rest of the items (dissuasion). Persons may be dissuaded from doing something but at the same time do not have to reconsider the thing from which they have been dissuaded. They may reconsider their future behavior – because they were dissuaded but may not reconsider the past behavior (which could potentially mean that they have understood/acknowledged what they did wrong (and that it was wrong) and therefore do not want to behave that way in the future). My suggestion would be to compare the conceptual definition of dissuasion with the present operationalization.
    - I agree – I've replaced that particular item with the following, which better aligns with the Focus Theory of Normative Conduct's argument that injunctive norms indicate what is considered appropriate vs. Inappropriate – „The response will make the first commenter believe their initial post was inappropriate.“
- Since the project proposes to use several measures that have not been validated before (e.g., toxicity dissuasion, willingness to contribute, etc.) it could be helpful to provide (post hoc) at least the evidence for unidimensionality from the CFA.
    - I've added a statement that we intend to conduct a CFA for each scale to provide evidence of unidimensionality in the Analysis Plan => Scale Reliability and Composites section.

- I was thinking about the justification of the smallest effect size of interest. Why exactly f = .11 is considered as being worthy of interest - (and potentially practically important)? What exactly does this size of the effect mean? Could unstandardized regression coefficients or means (from Pilot data) help to answer this question? - what is the magnitude of differences and are those differences high enough for being practically important?
  In our discussion of the SEOI (Proposed Experiment => Smallest Effect Size of Interest), we stated that the pilot mean differences of interest range from moderate (r = .33) to large (d =

1.25). Based on reviewer comments on the previous draft, we also included effect size estimates from relevant literature to provide converging evidence of the smallest effect size of interest, in case it was overestimated in our pilot sample. If the effect size does happen to be small (an r or f of .11), as it was in the Zerback and Fawzi (2017) paper, we thought it reasonable to use a sample large enough to find it.

- There is probably a typo on page 8: We plan to recruit 800 participants
- Thank you – this has been fixed.

Analyses plan

Original comment and reply italicized: 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.*
>
> > I see that you are planning to use mean substitution: „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." This is a very problematic technique, often leading to biased results and even the creators of Jamovi are stating on their blog that „Although there are a bunch of problems associated with mean substitution and you should probably never do it, it does make for a neat demonstration :P". There are many advanced techniques for the missing data imputation (some more some less difficult – multiple imputations, Full Information Maximum-Likelihood, random forests, Expectation maximization – but neither of them is available for Jamovi :(However, R can handle all of them.)
> >
> > Also, details as a percentage of questions that will be imputed („Those who answer a subset of questions") need to be specified in advance to decrease researchers degrees of freedom.
> >
> > I understand – I'll plan to use multiple imputation for missing values using the procedure described here: https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/
> >
> > I also specified in the manuscript that up to 50% of missing responses will be imputed: „Missing values will be imputed using multiple imputation for participants who answer at least 50% of the questions for a multi-item scale."

Original comment and reply italicized: 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.*
>
> > This is exactly my point, „reduction" and a „tendency to produce high-quality data" still do not rule out the possibility of having several careless respondents or the exclusion of non-careless respondents who just missed this one item (which could potentially bias the results). My suggestion was merely to choose a less conservative criterion to support the validity of the findings. Seven percent out of the 1122

participants means 78 potential careless responders who will be excluded based on this one attention check. To me, this seems to be a lot.

I appreciate this suggestion. I've added the additional attention check question below and specified in the manuscript that we will only drop participants who fail both attention checks.
"It is important for the quality of our data that we discard responses for individuals who aren't paying attention to the survey. To demonstrate that you are paying attention, please select "Green" for the following question.

What is your favorite color?"

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

Original comment and reply italicized: 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 may be wrong but the RR format should be designed for confirmatory studies, where the control of Type I error is the main goal, not for exploratory studies, which focus on the control of Type II error. This is the reason why authors of RRs should focus on minimizing researchers degrees of freedom – restraining any post hoc modifications to data analysis and interpretation. Therefore the method by which questions such as (How much should the conditions differ to conclude a difference? What does it mean „to a greater extent"? Any significant difference? Would that be theoretically/practically meaningful?) will be answered should be precisely defined in advance. Or H3: how much higher should the mean be? In H3 you have proposed the use of ANCOVA to determine whether means differ. To minimize researcher degrees of freedom, authors should also state how exactly will they do a comparison – based on what analysis will they conclude the difference. Based on planned comparison or post hoc test (what kind of test? e.g. Tukey, Holm, without correction…). The same also applies to H1 and H2.

Since this is exploratory, we don't have specific predictions as to the effect sizes so we unfortunately can't be more precise in that sense. However, we've clarified in the design table that we are looking for statistically significant differences: we have replaced "the two other conditions do not differ" with "the two other conditions do not differ significantly" and have also replaced the phrase "is similar to" with the phrase "does not differ significantly from" in all occurrences in the design table.

Since „to a greater extent" in H3 sounds clunky we've also replaced that phrase with the word „more." Again, we are not predicting a particular effect size, just any significant difference as specified in the design table.

I've specified in the design table which pairwise analyses, assuming an overall condition effect, will involve planned comparisons and which will involve post-hoc tests. I've also added to the design table how we intend to control the overall false discovery rate.

Original comment and reply italicized: 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.

Agree that the inclusion of another covariate would make it more complicated. At the same time keeping items that could compromise the validity does not help it either. I apologize if I have overlooked something but could it be possible to just substitute these three problematic items with neutral ones? (instead of doing a supplementary analysis with a smaller number of responses)? This would strengthen the validity of the study.

You're absolutely right. I've gone through and replaced all three pairs (4, 7, and 11) with examples that are less obviously about specific topics. Pair 3 in the pilot was rated as lower than I had thought in correcting (3.67, not 4) and higher than the others in going-along, so I replaced that pair, too. I've described these changes in the „Modifications to Pilot Procedure and Material" section in the manuscript (p. 28). I also noticed some mistakes in the word counts in Appendix A I was able to correct while making these edits. I think this is definitely a stronger stimulus set as a result – thank you.

I am very much looking forward to seeing the edited version of the manuscript and I wish the authors the best of luck with this work.

Best,
Marcel Martončik
Slovak Academy of Sciences